A Two-Stage Model for User's Examination Behavior in Mobile Search

Jiaxin Mao[†], Yiqun Liu[†*], Noriko Kando[‡], Zexue He[#], Min Zhang[†], Shaoping Ma[†]
[†]Department of Computer Science & Technology, Tsinghua University, Beijing, China
[‡]National Institute of Informatics, Tokyo, Japan
[#]Beijing Normal University
yiqunliu@tsinghua.edu.cn

ABSTRACT

With the rapid growth of mobile search, it is important to understand how users browse the mobile SERPs and allocate their limited attention to each result. To address this problem, we introduce a two-stage examination model that can separately capture the position bias with a skimming model and the attractiveness bias with an attractiveness model. The effectiveness of the proposed model is validated by using a dataset that contains explicit examination feedbacks from users. We further investigate user's examination behaviors by analyzing the model parameters learned via EM algorithm. The results reveal some interesting findings such as how the skimming behavior is dependent on the previous examination sequence and what factors are associated with the attractiveness of search results on mobile SERPs.

CCS CONCEPTS

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

KEYWORDS

Mobile Search, User Behavior Analysis, Examination, Selective Attention

1 INTRODUCTION

A good understanding of how users interact with the search engine may significantly contribute to improving its functionality. Among all kinds of user interactions, user's examination behavior on the search engine result page (SERP) draws much attention from the IR community. Many previous studies used eye-tracking to investigate examination behavior and found the *position bias* that user's attention are systematically biased towards the top-ranked results [5, 13]. Follow-up eye-tracking studies further showed that user's examination patterns are affected by a variety of factors such as the vertical types [11] and visual saliency of search results [9]. Combining with the *Examination Hypothesis* [13], which assumes that a user will click on a result on SERP when she has examined it and consider it as relevant or useful, these findings on examination behavior help us to accurately interpret the click-through data as implicit relevance feedbacks [5] and contribute to the improvement of result ranking.

With the rapid spreading of smartphones, understanding user's search behaviors on mobile phones become increasingly important. Because the user interfaces (UIs) of mobile devices are different from those of desktop computers, user's examination behaviors on these two platforms are different [6]. Knowing how users allocate their limited visual attention to mobile SERPs may be arguably more crucial because the mobile SERPs often contain information cards and knowledge graph results [7] that can provide users with sufficient information without requiring them to click. In these cases, examination is the only indicator of result usefulness and therefore an important feature to separate the *good abandonment* [14], where the information need is satisfied without any click, from the bad abandonment, where no relevant results are returned. Recently, Lagun et al. [7, 8] also used eye-tracking devices to inspect examination behaviors in mobile search. Besides characterizing the position bias in the mobile search environment, they showed that the browser viewport can be used as a measurement of user attention.

In this work, we want to further investigate and characterize user's examination behaviors in mobile search. Inspired by the previous work by Liu et al. [10], we propose a two-stage examination model for mobile search. Based on the assumption that the user will first *skim* a result and then decide whether or not to put more effort to *examine* the result based on its *attractiveness*, the two-stage model consists two components: a *skimming* model that captures user's browsing patterns on mobile SERPs and an *attractiveness* model that model the attractiveness of each search result. We use EM algorithm to fit the proposed models on a dataset collected in a carefully designed user study and investigate user examination behavior by analyzing the parameters of the fitted models.

Our study is different from the previous studies on this topic (e.g. [6–8]) in the following aspects: First, instead of using an eye-tracking device to record the fixation time on each result as a signal for examination, we collected user's *explicit examination* feedbacks. Liu et al.'s investigation in desktop settings [10] has shown that while the *examination* of the result is a necessary condition for click, the *skimming* event captured by eye-fixations does not always imply the examination or "reading" event measured by user's explicit examination feedbacks in retrospect. We assume that in mobile search, a thorough examination is

*Corresponding author

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

ACM ISBN 978-1-4503-4925-3/18/03…$15.00
https://doi.org/10.1145/3176349.3176891

necessary not only for making click decisions but also for processing the information on information card or knowledge graph results, so we adopt users' feedbacks as measures for examinations in our study. Second, rather than building a *discriminative* model to *estimate* examinations using features such as mouse movement for desktop search [10] and viewport time for mobile search [8], we construct a *generative* model, attempting to *explain* why some results are more likely to be examined by users. Because of these differences, our study is complementary to existing studies. We hope the findings in this study can shed new light on the understanding of user's examination behavior in mobile search.

2 A TWO-STAGE EXAMINATION MODEL

The basic assumption behind the two-stage examination model is that to examine a result on the SERP, the user will first *skim* the result and if the result is attractive to her, she will then put more effort to *examine* it. This assumption can be described in a more formal way:

$$(S_i = 1) \wedge (A_i = 1) \Leftrightarrow E_i = 1 \tag{1}$$

Here $S_i = 1$ means the user skimmed result i, $A_i = 1$ means result i is attractive, and $E_i = 1$ means the user examined result i.

We further assume that the probability of skimming $P(S_i)$ is determined by user's browsing patterns on mobile SERPs and independent of the content and appearance of the result and therefore independent of the probability of attractiveness $P(A_i)$. So the examination probability can be written as:

$$P(E_i = 1) = P(A_i = 1)P(S_i = 1) \tag{2}$$

We hope that the skimming model can model the position bias of examination while the attractiveness model can capture the attractiveness of the heterogeneous results in mobile search.

In this study, we use three different skimming models:

Rank model: $P(S_i = 1) = \gamma_{r_i}$, the skimming probability is determined by the rank of result i.

Position model: $P(S_i = 1) = \gamma_{y_i}$, the skimming probability is determined by the result i's Y position on the SERP. We use a binning method to estimate γ_{y_i}: each result is grouped into a bin according to its Y position on the SERP. The width of each bin is 160 pixels.

UBM model: The User Browsing Model (UBM) was originally introduced by Dupret and Piwowarski [3] to model user's examination and click behaviors in Web search. It assumes that the user will browse the SERP in a top-down order and the probability of examining a result depends on the rank of the results and the distance with the last previous clicked result. We use it to model the skimming probability as the following: $P(S_i = 1) = \gamma_{r_i, d_i}$, where r_i is the rank of result i, d_i is the distance in rank between result i and the last examined result.

For the attractiveness model that gives $P(A_i)$ for each result independently, we use a logistic regression model that maps a set of features x_i to $P(A_i)$:

$$P(A_i = 1) = \frac{1}{1 + exp(-x_i \cdot \beta)} \tag{3}$$

The features used to build the attractiveness model are shown in Table 1. We will further analyze how these features affect the result attractiveness and the overall performance of the two-stage model in Section 3.

Table 1: Features used in the attractiveness model.

Groups	Features	Descriptions
Content	height	The height of the result (in pixels).
	char_length	The length of the text content of the result, measured in number of characters.
	hl_length	The number of highlighted characters in the result.
	anchor_num	The number of hyperlink anchors in the result.
	image_num	The number of images in the result.
Visual	visual_saliency	The average/ sum/ standard deviation of the visual saliency map [4] of the result.
	edge_density	The average/ sum/ standard deviation of the edge density [2] of the result.
Annotation	relevance	The 4-level relevance annotation of the result.
	click_necessity	The 3-level click-necessity annotation of the result

We treat E_i as observable variables while S_i and A_i as latent variables in the two-stage examination model because only the explicit feedback for examination (E_i) was collected in the user study (See Section 3.1 for more details). Because of the existence of the latent variables, we use EM algorithm to fit the model [1]. For the attractiveness model, we estimate its parameters β in the M-step by using the logistic regression solver provided in `scikit-learn` [2] package.

3 EXPERIMENTS

3.1 Data Collection

We conducted a user study to collect a dataset [3] that contains participants' explicit examination feedbacks in mobile search.

The user study involves 20 search tasks and 43 participants. Each search task is defined by a query sampled from the query log of a commercial mobile search engine in China [4]. We wrote a background story for each query to create a simulated work task situation [1] for the participants. We also use each query to crawl four SERPs from four popular mobile search engines in China. Because the search tasks cover a wide range of topics, the crawled search results cover a variety of vertical types such as Image, Video, and Knowledge Graph. All the participants are college students aged from 19 to 23. 20 of them are female and 23 are male. All the participants are native Chinese speakers and reported that they were familiar with search engines and smartphones.

In the user study, we required each participant to use an Android Smartphone with a 5-inch, 1280×720, touchscreen to complete the 20 search tasks. For each search task, our experiment system would show one of the four crawled SERPs to the participants. While no further query reformulation is allowed in the user study, the participant could freely browse the SERP and click the results on them until she thought that she had completed the search tasks. The task order and the origins of the SERP were rotated to balance the impressions for each SERP and prevent potential order effect on participants' search behaviors. After the participant completed each search task, the experiment system would show the SERP again and asked the participant to provide binary explicit examination feedbacks ($E_i \in \{0, 1\}$) for all the results on the SERP and 4-level usefulness feedbacks for the examined results. The method of getting examination feedback is

[1] The derivation of the EM algorithm is similar to UBM model. We omit it here because of the limited space.
[2] http://scikit-learn.org/
[3] The dataset will be open to public after the reviewing.
[4] All the search tasks, instructions, and apparatus in the user study are in Chinese.

similar to that adopted by Liu et al. [10] in desktop settings. We acknowledged the limitation that the explicit examination feedbacks may be further affected by the position bias in the feedback process. But we chose not to randomize the result order in the feedback process because it would make the participant more likely to forget which results were actually examined by her during the search process.

After collecting the user behavior dataset, we also asked professional assessors to make 4-level relevance annotations and 3-level click necessity annotations [12] for the search results because we want to use them as features for the attractiveness model to inspect their relationship with the result attractiveness and examination probability.

In this way, we collected 860 search sessions. 919 distinct search results were shown 10,021 times in these search sessions. 2,765 result impressions were annotated as examined ($E_i = 1$). On average, 3.215 results were examined by the participants in each session.

3.2 Examination Prediction

Based on the collected data, we test whether the proposed two-stage examination model can effectively model the examination probability of mobile search results. To measure the performance in examination prediction, we use a 10-fold cross validation on our dataset to compute the log-likelihoods and perplexities of different examination models.

We measure the performance of the two-stage models that combine both the skimming models and attractiveness models. To show the advantage of building a two-stage generative model to separately model the position bias and attractiveness bias, we use a discriminative logistic regression model (LR) as baseline. The features used to build the baseline are the attractiveness features in Table 1 along with the the rank (r) and Y position (y) of the results. To test whether adding attractiveness models can improve the examination prediction performance, we further compare the log-likelihood and perplexity of the two-stage model against the corresponding skimming model.

The results are shown in in Table 2. We can see that: 1) except for the Position+Attr. Model, the other two models with different feature combinations outperform the logistic regression baselines, demonstrating the proposed generative models can better explain user's examination behaviors. 2) except for only using the visual features to build the attractiveness model, adding the attractiveness model as a component in the two-stage examination model significantly improves its performance.

3.3 Model Analysis

By analyzing the parameters of the fitted skimming and attractiveness models, we can characterize the position bias on examination in mobile search in different aspects.

For the skimming models, we show the γ parameters in the Position models and UBM models in Figure 1 and Figure 2. From Figure 1, we spot a sharp decreasing in γ_y within the initial viewport (first 4 bins, [0, 640 pix]), suggesting that in mobile search the position bias within the initial viewport is stronger. From Figure 2, we can see how the skimming probability is conditioned on previous examinations. The darker cells in the first column ($d = 1$) indicate that the user is more likely to skim a result that is right

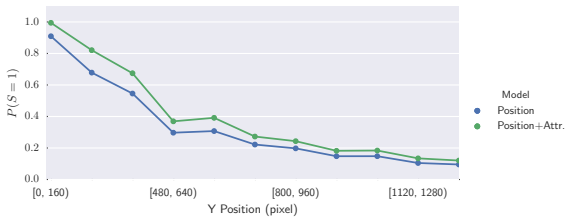

Figure 1: The skimming probability parameters γ_y estimated by Position and Position+Attr. models. The width of each bin is

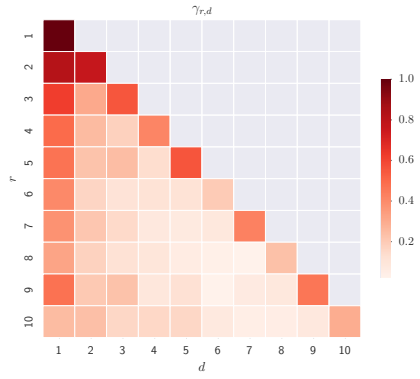

Figure 2: The skimming probability parameters $\gamma_{r,d}$ estimated by UBM+Attr. model.

below an examined result, while the darker cells along the diagonal ($d = r$) suggest that the user is more likely to skim a result if she has not examined any result yet.

For the attractiveness models, we show the β parameters of the UBM+Attr. models in Table 3. From this table, we can see that: 1) As expected, the height, character length, and highlighted character length is positively correlated with result attractiveness; 2) Both the visual saliency feature and the edge density feature have a positive correlation with result attractiveness. Users are more likely to examine a result with higher visual saliency. 3) Relevance is positively correlated with result attractiveness, suggesting relevance has an influence on the decision of putting more effort to examine the result or not. 4) Click necessity annotation is negatively correlated with attractiveness. This indicates that the results with low click necessities (e.g. knowledge graph and instant answer results) can indeed attract more attention from users. The top 3 most attractive and unattractive results according to $P(A_i = 1)$ computed by the attractiveness model of UBM+Attr. model are shown in Figure 3. The top 3 most attractive results are all federated results that consist of information from different sources while the top 3 most unattractive results are all query suggestions with four related queries.

4 CONCLUSIONS

To conclude, in this work, we introduce a two-stage examination model in mobile search. Using a dataset with explicit examination feedbacks from users, we show that the proposed model can

Table 2: The performance of the two-stage examination models measured in log-likelihood (*LL*) and perplexity (*Perplexity*). */ indicates the performance is significantly different from the baseline (LR) at $p < 0.05/0.01$ level. +/++ indicates the of the two-stage model is significantly different from the corresponding skimming model at $p < 0.05/0.01$ level.**

Feature groups:	Skimming model		Content		Visual		Content+Visual		Content+Visual+Annotation	
Eval. Metric:	*LL*	*Perplexity*	*LL*	*Perplexity*	*LL*	*Perplexity*	*LL*	*Perplexity*	*LL*	*Perplexity*
LR	-	-	-4.555	1.595	-4.547	1.594	-4.522	1.590	-4.490	1.584
Rank	-4.395	1.572	-4.369(**/++)	1.567(**/++)	-4.383(**/-)	1.569(**/-)	-4.361(**/+)	1.565(**/++)	-4.332(**/++)	1.560(**/++)
Position	-4.492	1.589	-4.481(*/+)	1.586(-/++)	-4.482(*/-)	1.587(-/-)	-4.472(-/-)	1.585(-/+)	-4.447(-/++)	1.580(-/++)
UBM	-4.183	1.539	-4.161(**/+)	1.535(**/++)	-4.172(**/-)	1.537(**/-)	-4.153(**/+)	1.533(**/++)	-4.122(**/++)	1.528(**/++)

Table 3: The parameters (normalized β) of the attractiveness model of UBM+Attr. model. We omit the the parameters that are not significantly different from zero at $p < 0.01$ level with '-'.

β	Content	Content+Visual	All
intercept.	1.834	0.077	-0.503
height	1.460	0.787	0.701
char_length	1.742	1.600	1.192
hl_length	2.049	1.718	0.378
anchor_num	-0.559	-0.242	-0.409
image_num	-2.661	-2.528	-1.387
avg. visual_saliency		0.526	0.412
sum. visual_saliency		0.883	0.688
std. visual_saliency		-0.564	-0.095
avg. edge_density		-	0.836
sum. edge_density		1.063	0.931
std. edge_density		1.185	0.628
relevance			1.605
click_necessity			-0.530

(a) Top 3 attractive results

(b) Top 3 unattractive results

Figure 3: Top 3 most attractive/unattractive results according to the attractiveness model of UBM+Attr. model.

effectively estimate the examination probability of each search result by separately capturing the position bias and attractiveness bias. We further analyze the parameters of the fitted models to characterize user's examination behaviors in mobile environment in different aspects such as how the skimming is conditioned on previous examinations and what features are associate result attractiveness in mobile search

We acknowledge the limitation of this study that we only use the explicit examination feedbacks from the participants in a small scale laboratory user study. For the future work, we can: 1) utilize eye-tracking device to investigate user's skimming (short fixation time on a result) and examination (long fixation time, reading sequence) behaviors; 2) collect a larger dataset that use remotely collected viewport data as signals for skimming and examination.

ACKNOWLEDGMENTS

This work was supported by Tsinghua University Initiative Scientific Research Program(2014Z21032), National Key Basic Research Program (2015CB358700), Natural Science Foundation (61532011, 61472206) of China and Tsinghua-Samsung Joint Laboratory for Intelligent Media Computing.

REFERENCES

[1] Pia Borlund. 2000. Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation* 56, 1 (2000), 71–90.
[2] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.
[3] 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.
[4] Jonathan Harel, Christof Koch, and Pietro Perona. 2007. Graph-based visual saliency. In *Advances in neural information processing systems*. 545–552.
[5] 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.
[6] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2016. Understanding eye movements on mobile devices for better presentation of search results. *JASIST* 67, 11 (2016), 2607–2619.
[7] 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.
[8] Dmitry Lagun, Donal McMahon, and Vidhya Navalpakkam. 2016. Understanding mobile searcher attention with rich ad formats. In *CIKM'16*. ACM, 599–608.
[9] Yiqun Liu, Zeyang Liu, Ke Zhou, Meng Wang, Huanbo Luan, Chao Wang, Min Zhang, and Shaoping Ma. 2016. Predicting search user examination with visual saliency. In *SIGIR'16*. ACM, 619–628.
[10] 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, 849–858.
[11] 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.
[12] 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.
[13] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting clicks: estimating the click-through rate for new ads. In *WWW'07*. ACM, 521–530.
[14] 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.