Investigating Examination Behavior of Image Search Users

Xiaohui Xie
DCST, Tsinghua University
xiexh_thu@163.com

Yiqun Liu
DCST, Tsinghua University
yiqunliu@tsinghua.edu.cn

Xiaochuan Wang
Sogou Incorporation
wxc@sogou-inc.com

Meng Wang
Hefei Institute of Technology
eric.mengwang@gmail.com

Zhijing Wu
DCST, Tsinghua University
wzjingzai@163.com

Yingying Wu
Department of Mathematics,
The University of Texas at Austin
ywu@math.utexas.edu

Min Zhang
DCST, Tsinghua University
z-m@tsinghua.edu.cn

Shaoping Ma
DCST, Tsinghua University
msp@tsinghua.edu.cn

ABSTRACT

Image search engines show results differently from general Web search engines in three key ways: (1) most Web-based image search engines adopt the two-dimensional result placement instead of the linear result list; (2) image searches show snapshots instead of snippets (query-dependent abstracts of landing pages) on search engine result pages (SERPs); and (3) pagination is usually not (explicitly) supported on image search SERPs, and users can view results without having to click on the "next page" button. Compared with the extensive study of user behavior in general Web search scenarios, there exists no thorough investigation how the different interaction mechanism of image search engines affects users' examination behavior. To shed light on this research question, we conducted an eye-tracking study to investigate users' examination behavior in image searches. We focus on the impacts of factors in examination including position, visual saliency, edge density, the existence of textual information, and human faces in result images. Three interesting findings indicate users' behavior biases: (1) instead of the traditional "Golden Triangle" phenomena in the user examination patterns of general Web search, we observe a middle-position bias, (2) besides the position factor, the content of image results (e.g., visual saliency) affects examination behavior, and (3) some popular behavior assumptions in general Web search (e.g., examination hypothesis) do not hold in image search scenarios. We predict users' examination behavior with different impact factors. Results show that combining position and visual content features can improve prediction in image searches.

CCS CONCEPTS

•**Information systems → Image search;** *Multimedia and multimodal retrieval;*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 ACM. 978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080799

KEYWORDS

Image search; Eye-tracking; Examination behavior;

1 INTRODUCTION

Multimedia content (e.g., images, video) has been increasingly incorporated into search engine result pages (SERPs) to increase both user experience and engagement. However, the way to show results in image search engines differs greatly from that of general Web search engines. Take the SERP in Figure 1 for example; in image searches, results are placed in a two-dimensional panel rather than a sequential result list. Instead of the query-dependent abstract of the landing page, the image snapshot is shown together with some meta information of the image (sometimes the meta information is only available with a hover behavior on the result), highlighted in Figure 1 with a red rectangle. Further, since results are joined, users can view results by scrolling up and down instead of clicking on the "next page" button.

User behavior data has been successfully adopted to improve general Web searches in result ranking [1, 36], query suggestion [7, 35], query auto completion [16, 17], etc. We therefore believe that understanding user interaction behavior in these multimedia search scenarios will also provide valuable insight into the optimization of their performances.

There exist a number of studies on user behavior log analysis of image search engines [2, 11, 20, 24, 30]. Click-through behaviors, query reformulation patterns, and session characteristics are investigated in [23, 24]. A comparison with general Web search behaviors was also performed in [2, 11]. Some researchers [15, 20, 22] focus on extracting implicit feedback signals from image search user behavior to improve result ranking performance. However, compared with the research on general Web searches, little attention has been paid to the examination behavior of image search users.

Examination is one of the prime concerns in existing search behavior studies because it is closely related with a user's attention distribution mechanism. With a better understanding of examination behavior, we can make better use of existing click-through signals, designing better evaluation metrics and helping users achieve their information needs more effectively. Although much research exists on search examination behavior, most efforts are focused on the SERP layout containing a single column of search results

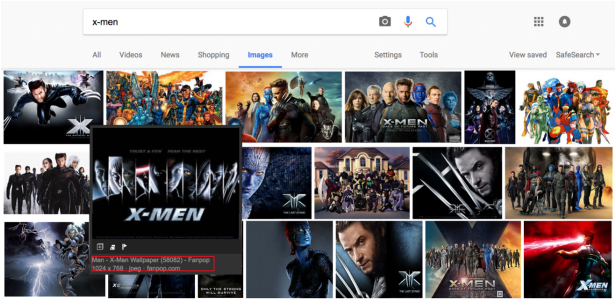

Figure 1: An example SERP in image search engine (The red box shows the meta information of the image while cursor hovering on the corresponding position)

[9, 19] or with some sidebar components such as knowledge graph [3, 34]. Considering that image SERPs usually present results in a two-dimensional placement style, whether the examination bias and hypothesis remain applicable becomes an open question.

As a preliminary attempt to model image search examination behavior, we perform a lab-based user study with the help of eye-tracking devices. We collect click-through, eye movement, mouse movement, and other behavior data during the search processes. The relevance scores of image results are also annotated to reveal its relationship with examination and click-through behaviors. We focus on the factors (behavior biases) that have impacts on user examination behavior and how to predict users' attention distribution using these factors. Specifically, this study addresses the following research questions:

- *RQ 1: How do positions of results affect user examination behaviors in image searches: are there also vertical position biases as in general Web searches? How do the horizontal positions of results affect user examination?*
- *RQ 2: How do the content of image search results affect user examination behaviors? Do content factors such as saliency, edge density, the existence of human faces, or textual information affect user examination during image searches?*
- *RQ 3: What are the relations between relevance, examination, and click in image searches? Does the examination hypothesis [8] still hold or there are new patterns?*
- *RQ 4: Can we predict examination behavior in image search? How can this prediction help us better understand users' examination behaviors?*

The paper is organized as follows. Section 2 reviews related work. Section 3 introduces our user-study settings. Sections 4-6 present findings from the user study. We report the experimental results of examination behavior prediction in Section 7. Finally, Section 8 discusses conclusions and future work.

2 RELATED WORK

We briefly summarize the related work in two categories: user behavior in image search and examination behavior of Web users. The former concentrates on analyzing large scale image search logs, and the latter employs eye-tracking devices to investigate user examination behavior in different settings.

2.1 User behavior in Image Search

With the high volume of traffic on Web search engines, the analysis of query logs becomes one of the most common approaches to understand user behavior. Previous works in image search also characterized the general user behaviors based on search logs [2, 11, 20, 24, 30]. Many features like query reformulation patterns, session length, and the number of viewed result pages are measured. Compared to general Web (text) search, image search leads to shorter queries, tends to be more exploratory, and requires more interactions. Park et al. [23] analyzed a large-scale query log from Yahoo Image Search to investigate user behavior toward different query types and identified important behavioral differences across them.

Interactive behaviors with image search result pages contain abundant implicit user feedback. Previous studies on multimedia search [14, 15, 22] explored user click-through data to bridge user intention gaps for image searches. O'Hare et al. [20] proposed a number of implicit relevance feedback features based on additional interactions including hover-through rate, converted-hover rate, and referral page click-through to improve image search result-ranking performance.

Most of the above works focused on mouse-based interactive behavior and tried to improve result ranking performance with corresponding features. Although Wang et al. [32] showed that understanding user examination behavior in Web searches provides powerful insight into user behavior modeling, little attention has been paid to examination behavior in multimedia searches.

2.2 Examination Behavior of Web Users

Understanding examination behavior is important for advertising, interaction design, and result ranking in Web search researches. Compared to techniques that rely on the explicit user actions (e.g., mouse clicks), eye-tracking yields more detailed moment-by-moment observations about how users interact with search information [13]. Cutrell et al. [9] used eye-tracking to explore the effects of changing in the presentation of search results. Pan et al. [21] and Buscher et al. [5] explored the determinants of ocular behavior on a single web page and tried to predict salient regions on Web pages. Tatler and Vincent [29] discovered that understanding eye-moving biases triggers gazing decisions in complex scenes. Underwood et al. [31] investigated how eye movements are affected by visual saliency and the semantic incongruence when inspecting pictures. The above works showed the effectiveness of eye-tracking in understanding examination behavior on images in different settings. In this study, we carry out an eye-tracking experiment to investigate the user examination behavior in image searches.

Click models also are used to model users' click and examination behavior in Web searches. Previous works on click models and search user's examination behavior patterns [8, 10, 32] were based on general search result pages with one-dimensional result lists. The major variables that most click models consider include examination, click, and relevance. They follow the examination hypothesis [8] that a clicked document should satisfy two independent conditions: it is examined and it is relevant. The cascade model [8] assumes that while a user examines the results from top to bottom sequentially, she/he immediately decides whether to click on a result. The partially observable Markov Model (POM) [33] treats

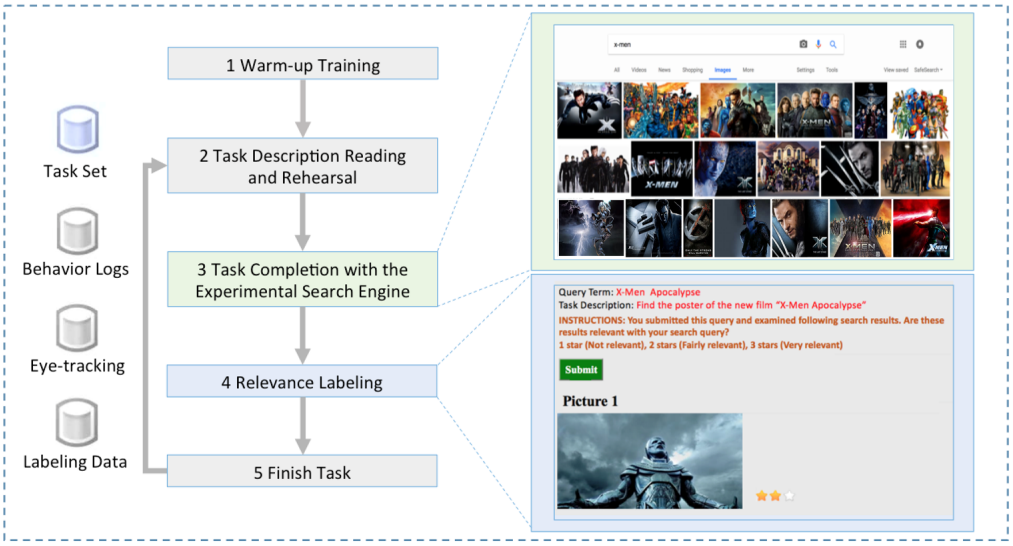

Figure 2: Data Collection Procedure

Table 1: Examples of Search Queries and Corresponding Taxonomies

Query Type [23]	#Query	Query Example (Task Description)
Specific	8	Doraemon (Find an image of Doraemon for your Wechat profile picture)
Generic	7 (1 for Warm-up Training)	New York (Find images about New York City's beautiful scenery)
Abstract	6	pleasant surprise (Find some images that express pleasant surprise to make a WeChat expression)

the user examination events as a partially observable stochastic process. Wang et al. [32] investigated the user's non-sequential examination behavior in ordinary search result pages and proposed a click model named PSCM (Partially Sequential Click Model) that captures this behavior. To the best of our knowledge, no existing click models are employed to model user examination behavior in image searches. Through our investigation in user examination behavior and its relation to clicks and relevance in two-dimensional result pages we may help design better click models that describes the interaction processes of image search users more precisely.

3 USER BEHAVIOR DATASET

3.1 Data Collection procedure

To conduct the tasks of our experiment , we sampled 21 intermediate frequency queries from the query log of a popular image commercial search engine[1]. Since search behaviors may be affected by different types of queries [23], we tried to design tasks to cover different query categories. We applied the Shatford-Panofsky approach [26] for image classification, like previous work in image search analysis has performed in [23]. In the sampled query set, eight queries are "Specific" (e.g., "Doraemon"), seven queries are "Generic" (e.g., "New York"), and six queries are "Abstract" (e.g., "pleasant surprise"). We also provided a task description for each query to make sure that users know exactly what results need to be looked for on the image SERPs. Table 1 shows a sample set of search queries and their corresponding descriptions. We crawled these queries' result pages from the commercial search engine and displayed them as SERPs in our experiment. The SERPs in the experiment contain 20 rows of image results to ensure a reasonable user study duration.

To investigate user's examination behavior during the search process, we carry out a laboratory study. The process is shown in Figure 2. Our user study consists of two stages. In the first stage, the participants perform several image search tasks, and meanwhile, their eye movement behaviors are collected. In the second stage, the participants are asked to annotate the relevance of each examined image search result. The dataset involves 46 undergraduate students majored in science, engineering and arts. All of them were frequent users of Web search engines. Because of calibration problems with the eye-tracking devices, not all participants' eye movement data were available, and 40 of them (female=16, male=24) were finally taken into account. Among the 40 participants, 20 performed the first stage of the experiment while the others took part in both stages, including recording eye-tracking data and labeling relevance.

At the beginning of the first stage, participants are given 21 search tasks, which include 1 warm-up task to familiarize them with the procedure and other 20 formal tasks. For each search task, participants were presented with an initial search query and a short description about the task. After reading the instructions, users click "start" button and a retrieved SERP containing 20 rows of pictures for this query is returned. Participants are instructed

[1]The data overs the period of June in the year of 2016, which we will release after the review process

Table 2: Relevance Labeling Dataset

#Images	#Images with same relevance judgement (66%)	
	#Relevant Images	#Irrelevant Images
1522	585	414

to examine the SERPs as they normally do. Like practical search scenario, they could scroll to move the page up and down, use the mouse to see hover text, and click a thumbnail to view and download the full-size image in the preview page. In the warm-up task, participants can browse and click the search results, and they can adjust the sensitivity of the mouse to the most appropriate level as well. We do not give any further instruction on which results to click on or when to end until participants are familiar with the experimental search engine.

The second stage is about relevance labeling. For each task, participants are given a relevance labeling page containing 20 images sampled from the images that they examined in the first stage in SERP. Users are asked to label each image with one of three levels (1-not relevant; 2-fairly relevant; 3-very relevant), consistent with previous works on query-image pairs relevance labeling[14, 37]. We deem the images to be relevant if their labels are score 2 or score 3. To further make sure the quality of the relevance judgment, we remove all images which participants have different opinions in their relevance and only retain the ones that different participants have the same opinions. The remaining annotated image set contains 999 images, in which 585 are relevant and 414 are irrelevant. Details about our relevance labeling dataset are shown in Table 2.

3.2 Data Collection System

In our study, we inject customized JavaScript into search result pages to log mouse activities on search pages when users perform search tasks. Considering that head-free eye trackers allow the collected interactions to be more natural and realistic, a Tobii X2-30 eye tracker is used to capture participants' eye movements. The search system is deployed on a 17-inch LCD monitor whose resolution is 1366×768. Google Chrome browser is used to display results of search system. To identify users' examination behavior, we detect fixations using built-in algorithms and all default parameters from Tobii Studio. All the collected data including the relevance judgments will be open to public after the double-blind review process.

4 POSITION BIAS AND EXAMINATION

With the user behavior data set described in Section 3, we look into the examination behavior of image search users and try to answer the four research questions in Section 1. Firstly, we focus on the influence of the position factor ($RQ1$).

4.1 Two-dimensional Placement of Results

As stated in previous sections, image search results are placed in a two-dimension manner. In the image search engine from which we collected SERPs, each result page contains 5 rows (row height = 200 px) and there are 3 to 7 images in each row (4 to 6 images in most cases). The users can scroll across different pages without having to click a "next page" button as in general Web search. Since we retain only the top 20 rows of results for each query, the SERP we collected each contains 4 "pages" of results (although the pages are only conceptional and there is only a page number shown on the SERP which is usually omitted by users).

Considering the fact that each row may contain different numbers of results (for example, the first row in the SERP from Figure 1 contains 6 images while the third contains 7), we decide to use absolute position instead of the border of images to segment the SERPs. Specifically, for each SERP, we equally divided it into 20 rows and 5 columns and use the median values of fixation duration/first arrival time in each grid to show users' examination patterns. Median is used instead of mean values here to reduce effects of outliers.

4.2 First arrival time

First arrival time of an image is the first time that a user gazes at the image. It is usually used to show user's examination sequence facing a given Web page. Figure 3(a) shows the heat map of the first arrival time of the images in the first result page, where the y-axis denotes the row number from 1 to 5. In Figure 3(b), we plot another heat map which represents the first arrival time of the images in page 1 to 4, with y-axis being the page number. The x-axis in Figure 3 indicates the column number as noted in Section 4.1. From Figure 3(a), we can see that the minimum first arrival time of the first row is at the position (1, 2), which is located around the middle position of the first row and highlighted by a black rectangular frame in the figure.

From both Figures 3(a) and 3(b), we can see that users generally follow a top-down pattern in the vertical direction of their examination sequences since the first row (in 3(a)) and the first page (in 3(b)) are both firstly fixated. However, in both of the first two rows in 3(a), the leftmost positions are not examined firstly. This is an interesting finding considering that the top-left position is usually regarded as the firstly-viewed position on a Web page [13]. In most current ranking strategies of image search engines, the most relevant image is usually placed at the top-left position (1, 1). However, according to our results, perhaps (1, 2) is a better choice since users pay attention firstly to this position (1, 2).

4.3 Examination duration

Examination duration of an image is defined as the dwell time during which a user examines the image. Dwell time has been regarded as an important implicit feedback for improving result ranking in search engines [21]. We calculate the examination duration for each image and plot users' examination duration distribution in Figure 4.

Similar with Figure 3, we also notice a middle position bias in the horizontal direction in both Figures 4(a) and 4(b). In 4(a), the position (1, 3) is fixated the most (as highlighted by a black rectangular frame), which is located around the middle position of the first row with a median fixation duration of 1.35 seconds. The left-most position (1,1) of the first row is fixated longer than the right-most position (1,5), but the duration is still shorter than the middle position (1,3). In 4(b), we also find that the middle position of the first page is fixated the most. In each page, it seems that the middle positions always draw more attention than the left-most or the right-most positions. This leads to a "T-shape" fixation distribution instead of the "F-shape" one (or "Golden Triangle") as described in most existing general Web search studies [13].

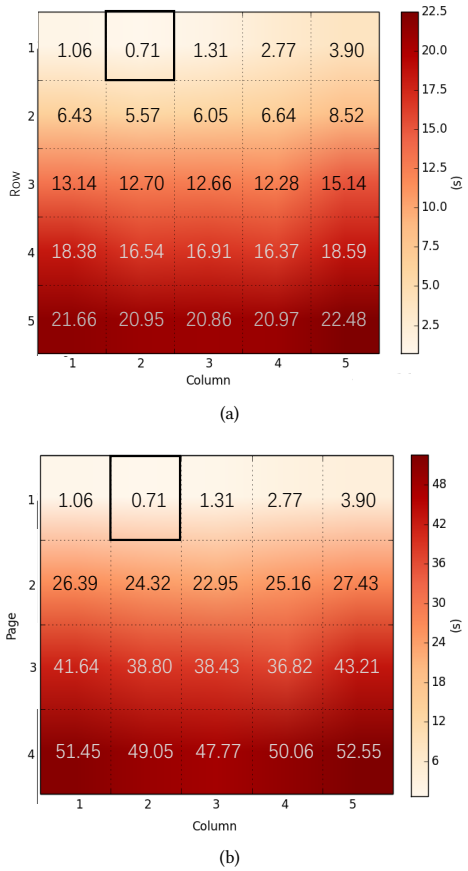

Figure 3: First arrival time distribution(s) in first page view (a) and global view (b)

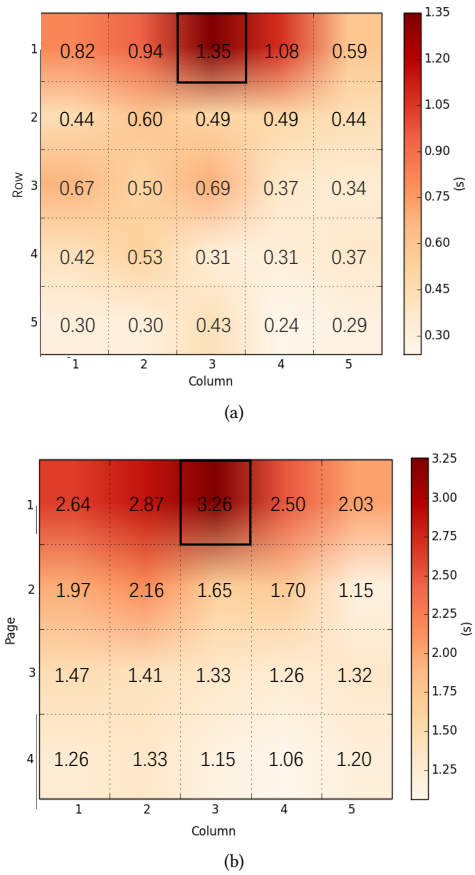

Figure 4: Examination duration distribution (s) in first page view (a) and global view(b)

In the vertical direction, we can see from Figures 4(a) and 4(b) that the first row and first page attract more user attention than the other positions. Generally, the fixation duration decreases with the row number or page number.

4.4 Statistical Modeling of Middle-position Bias

In this section, we statistically verify the observed middle-position bias in Figures 3 and 4 by analyzing the relation between the location of an image and gaze behaviors using a linear mixed model. The mixed model is a statistical model introduced by Bryk and Raudenbush [25]. The model contains both fixed effects and random effects, and they are particularly apt for multiple observations that are gathered over time on a set of people. The experiment was conducted with a set of 40 participants performing 20 search tasks, so we chose a linear mixed model, which is suitable for this data set size.

More specifically, we consider the effect of placing an image in the center columns (Columns 2, 3, and 4) on the first arrival time of the gaze. The dependent variable is "first arrival time," and the predictor variable is whether an image is placed in the center columns or on the sides (Columns 1 and 5). The random intercepts for both user and task are significant with $p < 2.2 \times 10^{-16}$ each. Therefore, we chose a linear mixed model with the user and task controlled as random effects. Testing the first arrival time with both the user and task controlled as random effects, the difference between the center columns and the sides is significant with $p < 2.2 \times 10^{-16}$. Hence the statistical model suggests that users tend to look first at pictures placed in the center columns and then at ones on the sides.

Similarly, we consider the effect on the fixation duration of gaze. The dependent variable is "fixation duration," and the predictor variable is whether an image is placed in the center columns or on the sides. The random intercepts for both user and task are significant with $p < 2.2 \times 10^{-16}$ each. Hence, we chose a linear mixed model with the user and task controlled as random effects as well. Testing the duration with both the user and task controlled as random effects, the difference is significant with $p < 2.2 \times 10^{-16}$. Therefore, the statistical model shows that the fixation duration tends to be longer for pictures placed in the center columns than for those on the sides.

In conclusion, eye gaze behaviors are related to the location of an image, and placing an image in the middle columns has a significant impact on both the first arrival time and the fixation

Table 3: Eye shift probability from images at Left column (a), middle column (b) and right column (c) to other images. Number 1 to 10 mean the distance (d) from image *i* to image *j* which defines according to the Equation 1. We only reserve the probability larger than 0.0001

(a)

	1	2	3	4	5	6	7	8	9	10	>10
Down	0.1137	0.0085	0.0006	0.0002	0.0001	0.0	0.0	0.0001	0.0	0.0	0.0044
Lower-right	0.0769	0.0693	0.0348	0.0101	0.0019	0.0	0.0001	0.0	0.0002	0.0	0.0334
Right	0.2187	0.1302	0.0685	0.0205	0.0043	0.0006	0.0	0.0	0.0	0.0	0.0001
Upper-right	0.0428	0.0458	0.0214	0.0093	0.0026	0.0005	0.0002	0.0001	0.0002	0.0001	0.0006
Up	0.0696	0.0060	0.0015	0.0010	0.0003	0.0003	0.0002	0.0	0.0003	0.0001	0.0001

(b)

	1	2	3	4	5	6	7	8	9	10	>10
Upper-left	0.0289	0.0150	0.0043	0.0012	0.0004	0.0004	0.0002	0.0001	0.0001	0.0002	0.0004
Left	0.1918	0.0497	0.0080	0.0013	0.0001	0.0	0.0	0.0	0.0	0.0	0.0
Lower-left	0.0567	0.0303	0.0074	0.0014	0.0002	0.0001	0.0	0.0	0.0001	0.0	0.0019
Down	0.0902	0.0128	0.0019	0.0004	0.0002	0.0	0.0001	0.0	0.0	0.0	0.0017
Lower-right	0.0518	0.0266	0.0060	0.0012	0.0001	0.0001	0.0	0.0	0.0	0.0	0.0021
Right	0.2143	0.0570	0.0095	0.0013	0.0001	0.0	0.0	0.0	0.0	0.0	0.0001
Upper-right	0.0296	0.0167	0.0049	0.0013	0.0003	0.0004	0.0002	0.0001	0.0001	0.0001	0.0003
Up	0.0541	0.0086	0.0024	0.0010	0.0007	0.0004	0.0003	0.0003	0.0002	0.0001	0.0003

(c)

	1	2	3	4	5	6	7	8	9	10	>10
Upper-left	0.0329	0.0337	0.0193	0.0066	0.0022	0.0008	0.0007	0.0001	0.0	0.0001	0.0009
Left	0.1966	0.1024	0.0662	0.0194	0.0030	0.0007	0.0	0.0	0.0	0.0	0.0
Lower-left	0.0978	0.1110	0.0641	0.0182	0.0052	0.0003	0.0001	0.0001	0.0	0.0	0.0039
Down	0.1139	0.0161	0.0016	0.0002	0.0001	0.0001	0.0	0.0	0.0	0.0001	0.0018
Up	0.0660	0.0082	0.0023	0.0004	0.0005	0.0009	0.0	0.0	0.0	0.0	0.0001

duration. These results agree with the observed middle-position bias in Section 4.2 and 4.3.

4.5 Transition Probabilities of Eye Fixation

Different from Web search results, the display of image search results is two-dimensional, so the transitions of fixation positions are in eight directions (upper left, up, upper right, right, lower right, down, lower left and left). To look into the eye movement patterns between different image search results, we show the transition probabilities for results in different directions and distances in Table 3. We also compare the differences of results in different positions (left-most column, right-most column and middle-column). Here, we define the distance (d) from image i to image j as:

$$d = max(abs[X_j - X_i], abs[Y_j - Y_i]) \quad (1)$$

Where X_i and X_j are row numbers of image i and image i while Y_i and Y_j are column numbers. The experimental data suggests the following findings.

- Most (44%) outgoing transitions from images located at the left column goes to the right while 23% goes lower right, 13% goes down, 12% goes upper right, and 8% goes up.
- Most (28%) outgoing transitions from images in the middle column goes to the right while 25% goes left, 11% goes down, 10% goes lower left, 10% goes lower right, 7% goes up, 5% goes upper right, and 4% goes upper left.
- Most (39%) outgoing transitions from images in the right column goes to the left, 30% goes lower left, 13% goes down, 10% goes upper left, and 8% goes up.
- In every direction, the transition probability decreases monotonically as the distance from the original image result increases.

From the transition probability, we infer that no matter where users' fixations are, eyes tend to move horizontally rather than vertically or diagonally. Results show a tendency toward a "near by principle" in which users prefer short-distance saccades than long-distance ones. These findings accord with existing researches in the mechanisms of eye movements [29].

5 IMAGE CONTENT AND EXAMINATION

Besides the position factor, we also want to investigate the influence of result content in examination. In many existing researchers, content of images have been shown to have impacts in people's attention allocation mechanisms. Therefore, we also want to look into the correlations between image content features and users' examination behavior including fixation duration and number of durations.

Table 4: Image content features

Feature Name	Feature Description
Saliency	Sum, mean and max of the given element.
Edge	Density of the given element.
Face	The number of human faces, the ratio of face areas.
Text	The number of texts, the ratio of text areas. (a Chinese character or an English word contributes for one text)

Table 5: Correlation between image content feature and examination behavior. *(or **) indicates the difference is significant at p < 0.05 (p < 0.01)

(coefficient, *p*-value)	Examination duration	Examination times
Sum Saliency	0.4842, **	0.5015, **
Mean Saliency	0.3442, **	0.3589, **
Max Saliency	0.2380,**	0.2583,**
Edge Density	0.1394,0.0532	0.1375,0.0566
Face num	0.0441,0.6731	0.0461,0.6589
The ratio of face areas	0.0324,0.7559	-0.0554,0.5956
Text num	0.1815,0.1889	0.0014,0.9919
The ratio of text areas	-0.0237,0.8648	-0.1801,0.1926

Image content features have been thoroughly surveyed in existing works. In [18], visual saliency features are demonstrated to significantly improve the success of examination prediction. Low-dimensional features like edge density are also found to be maximally informative features [4]. High-dimensional features like face and textual information are also widely used in image retrieval [27]. In this paper, we study four static image content features elaborated in Table 4. The GBVS algorithm [12] is applied to generate the visual saliency map. We use an edge detection algorithm introduced by Canny [6] to determine the boundaries of items (e.g., buildings) in the images. One of the authors manually labeled whether an image contains human face or textual information and use the labeling results as the features as well.

Given a list of image, we can extract static features from each image based on the method above. Further, each image's fixation duration and number of fixations are also collected by the eye-tracking devices. We use the Spearman's correlation coefficient which is a nonparametric measure of rank correlation used to assess the relation between image content features and users' examination behavior. The results are shown in Table 5 which shows that saliency has the highest correlation with fixation behavior in the sum, mean, and max of a given element, followed by the edge density feature. However, face features and the textual features are not closely related to the examination behavior. The reason why the high-dimensional features of faces and text have less influence on users' fixation may be that the number of faces and amount of text contained in image search results are limited, and sometimes they are auxiliary. For example, most texts in keyword-based image search results are watermarks rather than meaningful information.

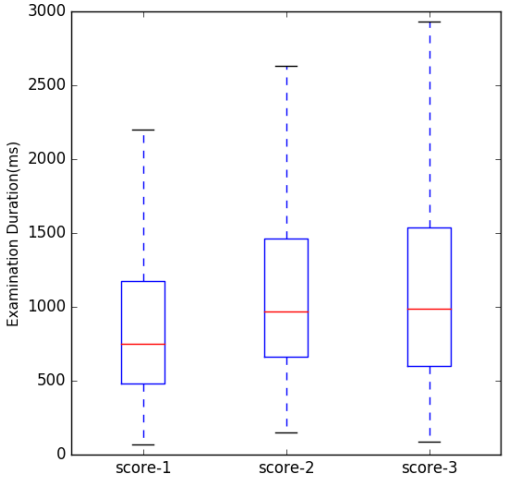

Figure 5: Examination duration boxplots for results with different relevance labeling scores

6 RELEVANCE, CLICK AND EXAMINATION

In the first stage of our experiment, we obtained participants' examination data and mouse behavior information including "click," "move," and "scroll" recorded by our built-in designed JavaScript. We conducted relevance labeling in the second stage. Thus, we obtained a feature space for selected images including examination duration, revisit behavior, click, and relevance score, which enabled us to investigate the relations between examination, click behavior, and relevance in our task. In this paper, We use pairwise T-test to verify the significance of differences between the different features and report the *p*-value.

6.1 Examination duration and Relevance

For each image in a task, we used a tuple with three elements (row, column, and duration) to describe it in an examination. An image has two or more tuples when there exist more than one fixations on it. We calculate the sum of examination duration for each image in the result panel if the image has two or more tuples. Based on the second stage of our experiment, we obtained a list of images with relevance scores. We draw a boxplot to illustrate the relation between examination duration and relevance in Figure 5, and perform a t-test between two of three image sets with different scores. Figure 5 shows that the mean examination duration of images with score 3 or score 2 is longer than images with score 1 (p-value < 0.01) while the difference between image set with score 2 and image set with score 3 is not significant. The above results show that when an image has a longer examination duration, it is more likely to be relevant, which may be labeled as "fairly relevant" or "very relevant". Thus, users' examination duration can be considered as a sign of image relevance, which implies that longer examination duration has a strong correlation with relevance.

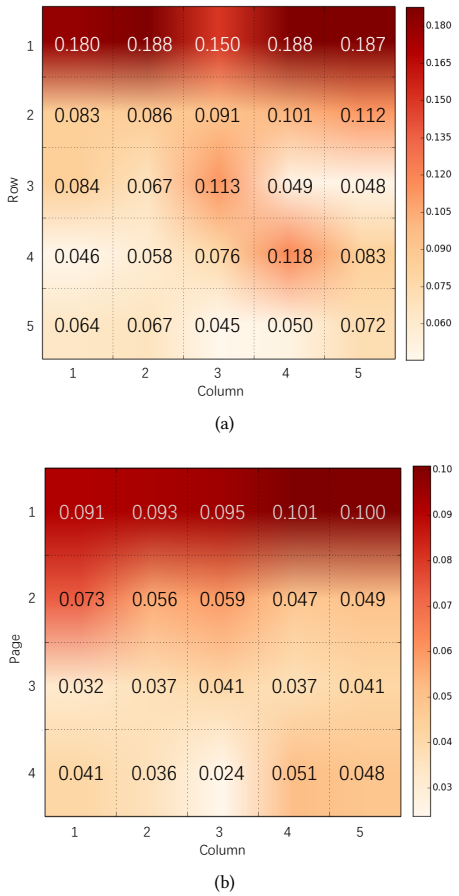

Figure 6: Click through rate distribution in first page view (a) and global view (b)

6.2 Revisit and Relevance

A revisit behavior in a task means that a user goes back to the previously examined image after examining another image. Here, we calculate the average scores of two groups of images (revisited and not revisited) The average score of revisited images in our task is 2.60 and higher than that of un-revisited ones (2.49). With paired two-tailed t-test, revisit behavior is significantly different between "relevant" and "irrelevant" (p-value < 0.05 for score 1 and score 2, and p-value < 0.01 for score 1 and score 3). This finding shares the same perspective in [36]. Xu et al. [36] shows that 50.2% of the revisited items in general web search have a high relevance.

6.3 Click and Relevance

Previous studies in image search [22, 28] indicate that image search click-through data is considerably more accurate in general than document-based search click-through data and can be used to boost the performance of an image search re-ranking system [15]. Therefore, based on the built-in JavaScript in our experimental platform, we investigate users' click behavior. Corresponding results are illustrated in Figure 6. Figure 6(a) shows the click-through rate for the first page. In contrast to the first arrival time and examination duration distribution, there is no noticeable middle-position bias on the first page. Furthermore, the four result pages overviewed in Figure 6(b) indicate that more clicks are observed on the first page and the left column of the second page, which is different from the middle position bias in examination behaviors shown in Figure 3(b) and Figure 4(b). Please be noted that if one of the images whose center positions are within the grid is clicked, then the grid receive one click.

The differences between click behavior and examination behavior (first arrival time, examination duration, and examination distribution) motivated us to further investigate the relation among click, examination, and relevance.

In this section, we address the two questions in $RQ3$: (1) Does examination hypothesis still hold in image search? (2) What is the relation between click, examination, and relevance? or will an image be clicked if it is examined and relevant with search target in image searches?

Examination hypothesis [8] assumes that a document being clicked ($C_i = 1$) accords with two conditions which are independent from each other: it is examined ($E_i = 1$) and it is relevant ($R_i = 1$). It is widely applied in Web search related studies such as click model constructions [10, 32]. Following this assumption, the probability of a document being clicked is determined as follows:

$$P(C_i = 1) = P(E_i = 1)P(R_i = 1) \tag{2}$$

According to the examination hypothesis, if an image in image searches is examined and relevant, then it will be clicked. We deem the images to be relevant if they obtain score 2 or score 3 and regard the images that have examination duration longer than 200ms as examined. If an image satisfies the examination hypothesis, it needs to satisfy any one of the following conditions: (1) it is relevant, examined and clicked; (2) it is not relevant and not clicked; (3) it is not examined and not clicked. Based on the experimental data, for each participant, we calculate the proportion of images that satisfy examination hypothesis. Result shows that 53% of images satisfy examination hypothesis while other 47% do not. Thus, in image search engine where results are shown in a completely different way with general Web search engines, examination hypothesis may not still hold.

To further investigate the relations among click, examination, and relevance, and to determine why the examination hypothesis is not applicable in image search, we calculated a confusion matrix that shows the relation between click (C) and relevance score, shown in Table 6. Because few images are not examined (6.7%), we only illustrate the results of examined images. We see that irrelevant images receive almost none click, which is consistent with the result in a general search. However, the probability of one relevant result being clicked is also very low ((4.08+4.84)%/(6.94+60.88)%=13.15%), which is not true according to Equation 2. This phenomenon may be caused by the fact that there are many similar images in the result panels in both the visual sense and content (for example, 4 out of 6 images in the first row of Figure 1 looks quite similar with each other). Sometimes there will be many relevant results. In contrast to general Web search results, image search results are self-contained, so that, users do not need to click the document as in a Web search to view the landing page in general. Instead, they can observe several images before deciding which ones to download or to click to see the larger version. Therefore, there is no need to

Table 6: Relation between click and relevance score given images are examined

E=1	C=1	C=0
Score=1	0.21%	26.72%
Score=2	4.08%	6.94%
Score=3	4.84%	60.88%

click each relevant image (many of them look quite similar with each other) to collect the required information and examination hypothesis doesn't hold any more here.

7 EXAMINATION BEHAVIOR PREDICTION

Through the eye-tracking study introduced in Sections 4-6, we obtained several interesting findings on user behavior biases including a middle-position bias in the user examination process and correlation biases between visual features and user examination behavior. As presented in Section 6, examination behavior can be an implicit feedback for content relevance. Therefore, predicting users' examination behavior based on static features that can be obtained offline is meaningful for search page optimization and UI evaluation.

7.1 Features and Model

As Table 5 illustrates, saliency and edge features play important roles in affecting the users' examination behavior. We extracted the visual saliency and edge density of images in our task into the feature space, which also contains image position (row, column) used to incorporate position bias into our prediction model. Because the first page receives more user examination as shown in Figure 4(b), the images in the first page are selected to construct our dataset which contains around 7,000 triples (participants,tasks and images). We set the input of the prediction model to be a combination of different features, and we compared the root mean square error (RMSE) of predictions by different feature combinations in Section 7.2. The prediction model outputs the probability of an image being examined, while the actual examination was obtained from the eye-tracking device as ground truth.

In this study, we compare different combinations of position and visual features. Our baseline feature group contains only the position feature (PF). Position bias is modeled in most existing click modeling efforts, so we use it as a reference to track how much improvement can be achieved by our method. Feature group SF involves only saliency features, and we use an index to mark which measurement (1-Sum, 2-Mean, 3-Max) it belongs. In feature group EF we only consider the edge density of images.

Given an image with a combination of features, our task is to predict its examination probability, which can be treated as a regression problem. In this paper, we apply a gradient boosting regression tree (GBRT) for the examination prediction tasks as in [18]. The prediction results of different combinations are compared with a 10-fold cross validation, and we report the average performance on the test folds.

Table 7: Examination prediction performance of different features groups

	RMSE	p-value
PF	0.1200	-
SF_1	0.1138	0.0904
PF+SF_1	0.1131	0.0279
PF+SF_2	0.1126	0.0296
PF+SF_3	0.1124	0.0070
PF+EF	0.1117	0.0332
PF+SF+EF	0.1098	0.0388

7.2 Prediction results and Discussions

We report RMSE for comparison, and a two-tailed t-test was performed to detect significant changes in the RMSE of prediction with position feature. Table 7 illustrates the corresponding results.

As Table 7 shows, adding visual saliency or edge information to a position-based feature (PF) can improve the RMSE of prediction. We find that PF+SF_1, PF+SF_2, PF+SF_3, and PF+EF significantly (with paired two-tailed t-test) outperform PF with p-value < 0.05 while PF+SF_3 has p-value < 0.01. Further, when we combine all features PF+SF+EF, the prediction model outperforms other combinations with p-value < 0.05. Thus, we can infer that combining position and visual features boosts the prediction performance in image searches. What is noteworthy is that as the features used in our prediction model are all static and can be calculated offline, they can be quite valuable for situations lacking users' behavior information.

8 CONCLUSIONS AND FUTURE WORK

In this paper, we carry out a lab-based user study with the help of eye-tracking devices in image search and obtain three interesting findings. (1) Based on first arrival time and examination duration analysis, We observe a middle-position bias of user's examination behavior, and we also apply a linear mixed model to justify the middle-position bias is significant statistically. Furthermore, we find that users are "lazy" in image search, they prefer small amplitude and short distance transitions between images. (2) Besides the position factor, visual features including saliency and edge density are demonstrated to have stronger correlation with user's examination behavior than high-dimensional features like the existences of human faces and textual features. (3) We investigate the relation between click, examination and relevance and find that examination duration is an useful implicit feedback for relevance of query-image pairs in image search. Also, as image results are self-contained, users do not need to click a document as in a Web search to view the landing page which results in the failure of examination hypothesis in image search scenarios.

Besides user study, we also perform an examination prediction experiment. Results show that combining position and visual features which can be calculated offline improves the prediction performance in image searches.

The research outputs of this paper are meaningful to inform the future image searches. For example, as there exists a middle-position bias of user's examination behavior, placing the most relevant results in the middle of each row rather than the leftmost positions may improve the users' satisfaction.

Our study makes a first step towards user's examination behavior analysis in image search. Interesting directions for future work include interaction behavior analysis result in preview pages (which is similar with but not completely the same as landing pages in general Web search). Moreover, we also plan to investigate user behavior in image searches on mobile devices and perform a comparison with PC based image searches.

ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation of China (Grant No. 61622208, 61532011, 61672311), Tsinghua University Initiative Scientific Research Program(2014Z21032), National Key Basic Research Program (2015CB358700). Wu was supported in part by NIH grant 5U54CA193313.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 19–26.

[2] Paul André, Edward Cutrell, Desney S Tan, and Greg Smith. 2009. Designing novel image search interfaces by understanding unique characteristics and usage. In *IFIP Conference on Human-Computer Interaction*. Springer, 340–353.

[3] Ioannis Arapakis and Luis A Leiva. 2016. Predicting user engagement with direct displays using mouse cursor information. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 599–608.

[4] Roland J Baddeley and Benjamin W Tatler. 2006. High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision research* 46, 18 (2006), 2824–2833.

[5] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. 2009. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 21–30.

[6] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.

[7] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. 2008. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 875–883.

[8] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 87–94.

[9] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 407–416.

[10] Georges E Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 331–338.

[11] Abby Goodrum and Amanda Spink. 1999. Visual Information Seeking: A Study of Image Queries on the World Wide Web.. In *Proceedings of the ASIS Annual Meeting*, Vol. 36. ERIC, 665–74.

[12] Jonathan Harel, Christof Koch, and Pietro Perona. 2006. Graph-based visual saliency. In *Advances in neural information processing systems*. 545–552.

[13] Gord Hotchkiss, Steve Alston, and Greg Edwards. 2005. Eye tracking study. *Research white paper, Enquiro Search Solutions Inc* (2005).

[14] Xian-Sheng Hua, Linjun Yang, Jingdong Wang, Jing Wang, Ming Ye, Kuansan Wang, Yong Rui, and Jin Li. 2013. Clickage: towards bridging semantic and intent gaps via mining click logs of search engines. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 243–252.

[15] Vidit Jain and Manik Varma. 2011. Learning to re-rank: query-dependent image re-ranking using click data. In *Proceedings of the 20th international conference on World wide web*. ACM, 277–286.

[16] Jyun-Yu Jiang, Yen-Yu Ke, Pao-Yu Chien, and Pu-Jen Cheng. 2014. Learning user reformulation behavior for query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 445–454.

[17] Yanen Li, Anlei Dong, Hongning Wang, Hongbo Deng, Yi Chang, and ChengXiang Zhai. 2014. A two-dimensional click model for query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 455–464.

[18] 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 *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 619–628.

[19] 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 *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 849–858.

[20] Neil O'Hare, Paloma de Juan, Rossano Schifanella, Yunlong He, Dawei Yin, and Yi Chang. 2016. Leveraging user interaction signals for web image search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 559–568.

[21] Bing Pan, Helene A Hembrooke, Geri K Gay, Laura A Granka, Matthew K Feusner, and Jill K Newman. 2004. The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*. ACM, 147–154.

[22] Yingwei Pan, Ting Yao, Tao Mei, Houqiang Li, Chong-Wah Ngo, and Yong Rui. 2014. Click-through-based cross-view learning for image search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 717–726.

[23] Jaimie Y Park, Neil O'Hare, Rossano Schifanella, Alejandro Jaimes, and Chin-Wan Chung. 2015. A large-scale study of user image search behavior on the web. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 985–994.

[24] Hsiao-Tieh Pu. 2005. A comparative analysis of web image and textual queries. *Online Information Review* 29, 5 (2005), 457–467.

[25] Stephen W Raudenbush and Anthony S Bryk. 2002. *Hierarchical linear models: Applications and data analysis methods*. Vol. 1.

[26] Sara Shatford. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly* 6, 3 (1986), 39–62.

[27] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence* 22, 12 (2000), 1349–1380.

[28] Gavin Smith and Helen Ashman. 2009. Evaluating implicit judgements from image search interactions. (2009).

[29] Benjamin W Tatler and Benjamin T Vincent. 2009. The prominence of behavioural biases in eye guidance. *Visual Cognition* 17, 6-7 (2009), 1029–1054.

[30] Dian Tjondronegoro, Amanda Spink, and Bernard J Jansen. 2009. A study and comparison of multimedia Web searching: 1997–2006. *Journal of the American Society for Information Science and Technology* 60, 9 (2009), 1756–1768.

[31] Geoffrey Underwood and Tom Foulsham. 2006. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly journal of experimental psychology* 59, 11 (2006), 1931–1949.

[32] Chao Wang, Yiqun Liu, Meng Wang, Ke Zhou, Jian-yun Nie, and Shaoping Ma. 2015. Incorporating non-sequential behavior into click models. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 283–292.

[33] Kuansan Wang, Nikolas Gloy, and Xiaolong Li. 2010. Inferring search behaviors using partially observable Markov (POM) model. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 211–220.

[34] Yue Wang, Dawei Yin, Luo Jie, Pengyuan Wang, Makoto Yamada, Yi Chang, and Qiaozhu Mei. 2016. Beyond ranking: Optimizing whole-page presentation. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 103–112.

[35] Wei Wu, Hang Li, and Jun Xu. 2013. Learning query and document similarities from click-through bipartite graph with metadata. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 687–696.

[36] Danqing Xu, Yiqun Liu, Min Zhang, Shaoping Ma, and Liyun Ru. 2012. Incorporating revisiting behaviors into click models. In *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 303–312.

[37] Yongdong Zhang, Xiaopeng Yang, and Tao Mei. 2014. Image search reranking with query-dependent click-based relevance feedback. *IEEE transactions on image processing* 23, 10 (2014), 4448–4459.