From Skimming to Reading: A Two-stage Examination Model for Web Search

Yiqun Liu[1], Chao Wang[1], Ke Zhou[2], Jian-Yun Nie[3], Min Zhang[1], Shaoping Ma[1]
[1]State Key Lab of Intelligent Technology & Systems, TNLIST, Department of Computer Science and Technology, Tsinghua University, [2]Yahoo Labs, London, U.K., [3]Université de Montréal
[1]{yiqunliu, z-m, msp}@tsinghua.edu.cn; [2]zhouke.nlp@gmail.com; [3]nie@iro.umontreal.ca

ABSTRACT
User's examination of search results is a key concept involved in all the click models. However, most studies assumed that eye fixation means examination and no further study has been carried out to better understand user's examination behavior. In this study, we design an experimental search engine to collect both the user's feedback on their examinations and the eye-tracking/click-through data. To our surprise, a large proportion (45.8%) of the results fixated by users are not recognized as being "read". Looking into the tracking data, we found that before the user actually "reads" the result, there is often a "skimming" step in which the user quickly looks at the result without reading it. We thus propose a two-stage examination model which composes of a first "from skimming to reading" stage (Stage 1) and a second "from reading to clicking" stage (Stage 2). We found that the biases (e.g. position bias, domain bias, attractiveness bias) considered in many studies impact in different ways in Stage 1 and Stage 2, which suggests that users make judgments according to different signals in different stages. We also show that the two-stage examination behaviors can be predicted with mouse movement behavior, which can be collected at large scale. Relevance estimation with the two-stage examination model also outperforms that with a single-stage examination model. This study shows that the user's examination of search results is a complex cognitive process that needs to be investigated in greater depth and this may have a significant impact on Web search.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Search process

Keywords
Eye tracking; Mouse movement; Selective attention; Web search

1. INTRODUCTION
Web search has reached a level at which a good understanding of user interactions may significantly impact its quality. Among all kinds of user interactions, examination is an important one that attracted much attention. Our understanding on how users allocate their limited attention to search engine result pages (SERPs) can contribute to improving search UI designing, result ranking, Ad delivery and many other research issues in Web search. It also plays a central role in the *Examination Hypothesis* [5, 27], which assumes that one result on SERP will be clicked only if it is examined.

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.
CIKM'14, November 3–7, 2014, Shanghai, China.
Copyright © 2014 ACM 978-1-4503-2598-1/14/11…$15.00.
http://dx.doi.org/10.1145/2661829.2661907

Many previous investigations on user examination bahavior relied on eye-tracking. Richardson [27] and Joachims [18] are among the first to point out that users are more likely to examine results near the top of SERPs based on findings in eye-tracking experiment. Cutrell [6] and Buscher [3] found that eye movements of users with different search intents are quite different. Wang [35] and Diaz [7] found that different result appearances may lead to different eye movement behaviors on both vertical and ordinary results. While all these studies based on eye-tracking have revealed a number of important findings in search users' examination process, they generally assumed that when a user fixes eye on a result for a certain time (e.g. above 200 ms), the result is examined. The eye fixation sequence was assumed to be that of examination, and the remaining eye movements were ignored. These studies follows the *Strong Eye-mind Hypothesis* [19], which supposes that what the eyes fixate on is what the mind processes. However, this strong assumption is not always validated. For example, Just *et al*. [19] found that while the duration of the gaze is closely related to the duration of cognitive processes, they are not necessarily identical. With a number of experiments, they found that the gaze duration may at best provide a rough estimate of the absolute duration of processing. Therefore, although the eye fixation sequence helps us understand users' examination patterns on SERPs, it may not necessarily reflect the true examination sequence of the user. In order to construct better models for user interactions (e.g. click models), we need to better understand the relationship between *fixation*, *reading* and *clicking*.

This study is an attempt to address the question. To this end, we design an experimental search engine system that collects simultaneously eye-tracking, mouse movement, click-through behavior and user's explicit feedback on result reading (see Section 3). We analyze in depth the examination process of Web search users and we find that the results fixated by users are different from those which users remember to have "read". It shows that user's examination process is non-trival and more complex than what eye fixation sequence shows. An example in Figure 1 shows a user's eye fixation sequences and the explicit feedback on reading during a search session. We can see that the user quickly looked through the first two results before focusing on the third one. Although all three results are fixated for some time, the user's explicit feedback showed that he/she only regarded the third result as being read. This typical example shows that the user's result reading process may not always be aligned with eye fixation. Therefore, simply treating fixation as examination (or use a eye fixation duration threshold as ground truth for result examination, see Section 4.1) may be misleading for the understanding of Web search user behaviors.

In addition, we also observe that before "reading", there are rich eye movements over the results (with vaiable fixation durations, or without eye fixation). They are important to understand how users examine the results, but are typically discarded in the previous studies. In this paper, these movements are considered to correspond to "skimming" – a quick overlook at the result without

reading. Reading a result requires that the result be first skimmed. Similarly, not all results read by the user are clicked by users while almost all clicked results are read by users. This observation motivates us to consider an examination as a two-stage process. In the first "skimming to reading" stage (Stage 1) which is featured by sometimes unconscious eye fixations, users quickly look through results and decide whether one result should be ignored or paid further attention to. In Figure 1, the user's attention on the first two results corresponds to Stage 1. In this example, the user chose to ignore these results. In the second "reading to clicking" stage (Stage 2) which is usually remembered by users, they carefully read and comprehend the results selected from Stage 1 and based on the reading, decide whether to click on it or not. In Figure 1, the user's examination on the third result might come into Stage 2, and the user remembers that it has been read.

Figure 1. A user's eye fixation sequence and the corresponding explicit feedback on result reading for top results in a search session of query "学雷锋作文" (*Essays on learning from Lei Feng* in Chinese). Radius of circle means fixation length.

Our proposed two-stage examination model and the choice of two stages are inspired by the attention selection mechanism [33] which is widely accepted in cognitive psychology studies. It says that human attention consists of two functionally independent, hierarchical stages: An early, pre-attentive stage (similar to Stage 1) that operates without capacity limitation and in parallel across the entire visual field, followed by a later, attentive limited-capacity stage (similar to Stage 2) that can deal with only one item (or at most a few items) at a time. Attention selection is one of the basic cognitive mechanisms of human beings and the two-stage examiantion model can be regarded as an attempt to explain how the mechanism works in Web search environment. What we propose in this paper is as follows:

- [**Two-stage Examination Model**] With analysis of user's search interaction process, we show that users may examine SERPs with a two-stage strategy. This two-stage examination model reveals the relationship among eye fixation, result reading and click-through behaviors. It also helps us to understand the mechanism with which search users allocate their attention selectively.
- [**Behavior Biases in Two-stage Examination**] While revisiting the search behavior biases including position bias [5], domain bias [13] and attractiveness bias [1, 22], we found that these biases have different impacts on user behavior in different examination stages. It means that users may rely on different signals to make decisions in different stages. These findings also reaffirm the necessity of the proposed two-stage model.
- [**Two-stage Examination and Relevance Prediction**] A prediction model is constructed to identify result examination in different stages with mouse movement information that could be collected at large scale. After that, a learning method is proposed to estimate the relevance of a result based on the two-stage examination model. The two-stage model is found to significantly outperform the orginal single-stage model.

The remainder of this paper is organized as follows. In Section 2 we review some related studies on user interactions in Web search. Section 3 describes the framework of the experimental system and the interaction data collected in our study. Section 4 analyzes the relationship between fixation, reading and click-through behaviors and proposes the two-stage examining model. Section 5 focuses on the behavior biases in the two-stage model. In Section 6 we attempt to predict two-stage examination behavior using mouse movement information and then use this information to estimate result relevance. In Section 7 we discuss the extension of our work before some concluding remarks.

2. RELATED WORK

Two lines of research are related to this work. One focuses on current endeavors to infer user intention and examination directly from the user's gaze movements on a SERP. Our work explores further in this line by looking into user's result reading process and we propose a two-stage examination model. The second line focuses on the relationship between gaze and mouse movement, and exploits mouse movement information for relevance prediction. We follow this line by utilizing mouse movement (rather than gaze) for relevance estimation using our two-stage examination model.

2.1 Eye-tracking Studies in Web Search

The application of eye-tracking devices to Web search has received a considerable amount of attention from both academia and industry. Eye-tracking devices allows researchers to record users' real-time eye movement information, which helps better understand how users examine results on SERPs.

Granka et al. [10], Richardson et al. [27] and Joachims et al. [18] use eye-tracking devices to analyze user's basic eye movements and sequence patterns throughout search tasks. Guan et al. [11] found that the decrease of user's attention in search sessions is closely related to query intents. Cutrell et al. [6] further investigated into how user's eye movement behavior varies for different query intents. Wang et al. [35] and Diaz et al. [7] found that different result appearances might create different biases on eye movement behavior for both vertical and other results on SERPs. Navalpakkam et al. [23] found that the flow of user attention on nonlinear page layouts is different from the widely believed top-down linear examination order of search results. Cole et al. [39] identify different user behavior patterns while performing different Web search tasks.

Based on these findings, a number of generative click models [4, 5, 8, 35] have been constructed to model users' behavior during the search process. Most of these studies follow the strong eye-mind hypothesis [19] and regard eye fixation sequences to be the same as user's examination sequences. However, cognitive processes may be more complex than what a simple eye fixation sequence can describe. Theeuwes et al. [34] showed that eyes will move to new objects unconsciously without the mind's control due to the selective attention mechanism [33]. Shiffrin et al. [30] pointed out that although overt attention (with eye fixation) is a significant part in cognitive processes, covert attention (usually without fixation) also helps to direct the gaze toward objects of interest. More importantly, Just et al. [33] found that there are no mapping rules between what is being fixated and what is being internally processed if the visual display is not relevant to the user's current task. Considering the many distracting factors on SERPs (e.g. ads, multimedia components and results that are not so relevant), it is difficult for us to assume that users always have full attention to all results. Therefore, whether strong eye-mind hypothesis holds in Web search remains to be further investigated.

The above studies show that we cannot simply regard eye fixations as the only sign of examination because the cognitive process in Web search is more complex than what the strong eye-mind hypothesis assumes. Different from most existing studies, we investigate the examination behavior by focusing on the relationship between eye fixation, result reading and click-through behaviors. Through these analyses, we hope to reveal the actual mechanism with which search users examine results on SERPs. To our best knowledge, there has been no previous work on this topic.

2.2 Eye-Mouse Coordination and Mouse movement studies in Web search

While eye movements during a search process could give us much insight into users' examination behavior, it is not useable at large-scale in practice. Therefore, many researchers tried to use instead mouse movement information, which could be collected at large scale, to simulate eye movements. Rodden [28] identified multiple patterns of eye-mouse coordination, including the mouse following behavior in both x and y directions while the eye inspected results. They also found a general correlation between eye and mouse position, where the centers of the distribution of the eye/mouse distances are quite close to each other. Huang et al. [15] extended these findings by investigating variations in eye-mouse distances over time. They found that the distance between eye fixated point and cursor peaked approximately 600 ms after page loading and decreased over time. They also found that the mouse tended to be behind eye gaze by approximately 700 ms on average.

Huang et al. [17] found correlations between result relevance and the cursor hovering behavior on the SERP. They incorporated mouse hover and scroll information as additional signals into click models to improve click prediction performance [16]. Guo et al. [12] analyzed the relationship between examination patterns and result relevance from post-click behaviors including cursor movements on landing pages. They constructed a predictive model to capture these patterns in order to improve search result ranking. As user's mouse movements on landing pages are also difficult to collect on commercial search engines, Speicher et al. [32] built a system to collect user's mouse movement information on SERPs and tried to predict result relevance using this information. The system showed a better relevance prediction performance than some existing click models on a search engine of hotel information. Smucker [38] focuses on mouse behavior during relevance judgment process and find that mouse behavior may not be a good sign for relevance.

In this work, we also build a laboratory system to collect user's examination behavior. We find that users may examine Web search results with a two-stage examination process and user's relevance judgments on these two examination stages are different. Taking this difference into account may be helpful to improve relevance prediction. Since we would like to employ our two-stage model for practical Web search applications and the previous work showed that mouse movement can be aligned with eye-gaze well, we also use mouse movement information to predict whether results are examined in different stages in this work. The examination prediction results are then adopted to estimate result relevance with examination hypothesis. Through this relevance prediction framework we want to show that the proposed two-stage examination model could better extract users' implicit relevance feedback information.

3. COLLECTING USER BEHAVIOR DATA

3.1 The Experimental Search Engine System

To analyze users' interaction behaviors on SERPs, we design and implement a lab-based search engine to collect user behavior data. From the experiment process shown in Figure 2, we can see that it can collect four types of user behavior information for each search task: (1) eye movements, (2) mouse movements, (3) click-through information and (4) users' explicit feedback on result reading.

As shown in Figure 2, the process of this study is as follows. Firstly, we prepare a set of search tasks and their corresponding fixed queries (one query for each task). To make sure that the same SERP for a certain task is shown to all the participants in the experiment, we crawled and stored in advance the corresponding SERPs of all search tasks. Since multimedia components on SERPs may influence user's eye movements and click-through behavior [35], which is beyond the scope of this study, we removed advertisements and vertical results so that each SERP contains exactly 10 organic results. The results are shown on the same screen whose resolution is 1920*1080 for each participant.

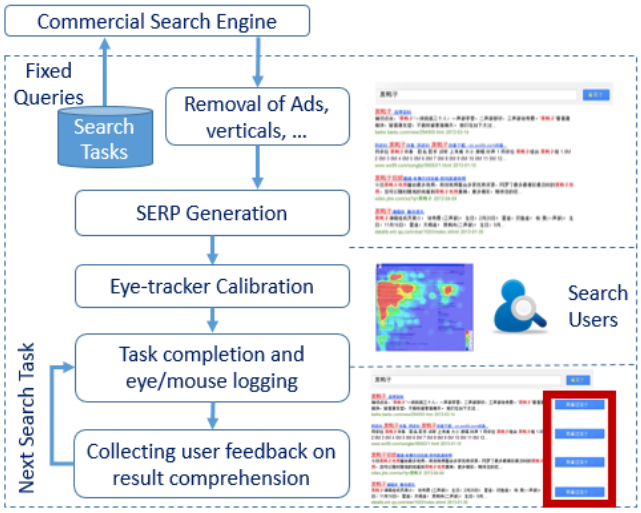

Figure 2. The experimental system for collecting eye/mouse movement/click-through data as well as explicit feedback information of result reading on SERPs

To collect reliable eye movement information, before the experiment, each participant should first go through a calibration process as required by the eye tracker. We used a Tobii X2-30 eye-tracker with its default parameter setting.

The participants are instructed to finish a number of search tasks with the experimental search engine. During the search process for each task, their eye movements were recorded by the eye-tracker and their mouse movements/click-through behaviors were also logged by injected JavaScript code on SERPs.

Right after the search process of each task, participants were required to label each of the results as "read" or "not read" before moving to the next task. To ensure the quality of feedback information in this step, we use two example search queries (one navigational query and one informational query) to show the participants how to finish the labeling task before they actually start the experiment. An instruction card is also given to each participant so that he/she could refer to the annotation rules anytime during the search process. In the instructions to participants, the process of labeling "result reading" is described as "*please label whether you have read and comprehended one result (according to its title, snippets, url, etc.) during this search session by clicking on the button shown beside it*". The buttons which are adopted to record this explicit feedback information are marked by the red box in Figure 2. To avoid affecting users with additional information during the feedback process, we do not change the color of the clicked hyperlinks as most navigators do.

We believe that understanding the reading behavior in Web search scenario is a challenging task on its own. The cognitive process of Web search user is rather personal and the perceiving of document relevance is only possible for the user himself/herself. In addition to collecting eye gaze and mouse movements, we believe that collecting explicit feedback from the user is the best we can do (we will show the effectiveness of this in our experimental results).

We should note that the definition of "reading" in the explicit feedback on "read" or "unread" results (as well as in the rest of the paper) is not exact the same as the definition of "text reading" in some cognitive studies such as [25]. Our definition is broader, meaning that the user has understood at least some of the result's content through its title, snippet and url. Meanwhile, "reading" in most cognitive studies focus on the process of decoding symbols to derive meaning from text. In other words, we focus on the outcome of search result-level reading instead of word-level or phrase-level decoding processes.

3.2 Participants and Search Tasks

Altogether 37 participants (21 males and 16 females, with a variety of self-reported Web search expertise) were recruited for this lab study. The subjects are all undergraduate students from a Chinese university in their first year whose majors include engineering, journalism, biology and law. The number of subjects is similar to other search eye-tracking studies [6, 10].

25 search tasks sampled from Sogou.com (China's second largest search engine) click-through logs [1] were assigned to each participant in the experiment. Each task was specified by a fixed median-frequency query together with a description of the information need to avoid ambiguity. As different types of query intent [2] may lead to different examination behaviors [10], we retain 5 navigational queries and 20 informational/transactional queries in the query set. This distribution roughly follows the proportions of the task types in Web search [2]. The search tasks are assigned to the participants in the same order with navigational ones randomly mixed with others.

With the experimental system, we could record each participant's eye/mouse movement information on each result for each search task. Behavior data from several query sessions were removed due to participants' operation errors or software crashes. In total we collected 8,900 valid <*user*, *query*, *result*> tuples[2]. For each tuple, the corresponding fixation behavior, click-through behavior and explicit feedback information on result reading are recorded by the system. Therefore, we could investigate the relationship between fixation, reading and click-through behavior with this dataset.

3.3 Relevance Annotation of Search Results

To evaluate the relevance judgment performance, relevance scores of all search results (25 queries, 10 results for each query) were explicitly labeled using a four point scale ranging from "Good", "Fair", "Poor" to "Bad" in diminishing order. In the following experiments, results with labels "Good" and "Fair" are considered as "relevant", and results with other labels are treated as "non-relevant". Three professional assessors from a commercial search engine company annotated all the results on their own and the Kappa coefficient [31] of "relevant" and "non-relevant" among assessors is 0.727, which means agreement at a substantial confidence level. We use majority voting to combine the annotation results (binary judgments) from assessors as the ground truth.

[1] Sogou search log sample: http://www.sogou.com/labs/dl/q-e.html.

4. TWO-STAGE EXAMINATION MODEL

4.1 From Skimming to Reading

Using the data set described in Section 3, we try to find out whether fixation is equal to reading as strong eye-mind hypothesis and most existing search-based eye-tracking studies assume. Table 1 shows the percentages of <*user*, *query*, *result*> tuples according to whether they were fixated and whether they were annotated as "read". For the threshold of fixation, we adopt the same practice as most previous works (200-500 milliseconds as in [21, 29]) and set it to 500 milliseconds. We also tried a number of other thresholds varying from 250ms to 2000ms and got similar results. Readers can refer to our eye-tracking data set for details.

From the results in Table 1 we can see that the majority of tuples follow the strong eye-mind hypothesis, as fixation and reading are identical for 65.70% of them. However, there are also 34.30% tuples with different fixation and reading values. It means that these tuples do not follow the strong eye-mind hypothesis. Considering the fact that 45.80% tuples with fixation=1 are not annotated by users as "read", we may not simply treat fixation as the only sign for examination. This motivates us to further investigate the user's reading process during the examination.

Table 1. Distribution of <*user*, *query*, *result*> tuples with respect to fixation and reading behaviors

	Fixation=0	Fixation=1
Reading=0	31.61%	28.81%
Reading=1	5.49%	34.09%

There are 5.49% tuples with fixation being 0 (user fixated on the corresponding result for less than 500ms or did not fixate on it) but reading being 1. We believe that this small proportion of results may be due to the existence of covert attention as discussed in [30], which means that users may also acquire some information without fixation. Memory confusion or eye-tracking device errors may also be possible reasons. However, the small number indicates that most of our experimental data is reliable. Regardless of these possible noises in data, we find that most (3034 out of 3523) of the "read" tuples are with fixations. Therefore, we propose the following hypothesis:

Hypothesis 1 (Fixation Hypothesis): *Eye fixation on a search result is a prerequisite for reading this result.*

We also looked at the fixation duration of the results (see Figure 3) whose fixation is 1. It is found that the fixation of "read" results is significantly longer than that of "unread" ones (2686ms vs. 1962ms on average, with two-tailed t-test $p<0.001$). It shows that the "read" results are paid more attention to by users while many of the "unread" ones are just skimmed without careful comprehension. This may also explain why a user does not always examine all results fixated on, because the careful reading process requires much longer period of time.

From Figure 3 we can also see that although the average fixation length of "read" results is longer, there are also several "unread" results whose fixation lengths are over 2000 or even 4000 milliseconds. When we asked about the reasons to the participants who annotated this kind of "long-fixated" results as unread, most of them said that when they fixated on the results, they were actually distracted and were think about something else (related knowledge, the task requirement, etc.). This shows that it is inappropriate to assume that each fixation means examination.

[2] The data set is open to public to promote reproducibility: http://www.thuir.cn/group/~yqliu/publications/cikm2014-liu.7z.

Adopting a large fixation length as the threshold for result examination may reduce noises, but as we can see from this figure, the relationship between reading behavior and fixation length are much more complex than simple threshold settings. Simply using one fixed threshold on fixation length could not capture accurate reading behavior.

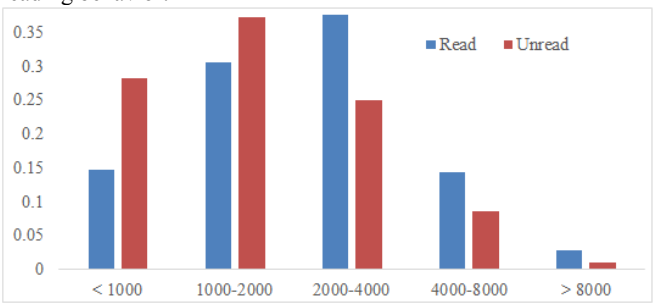

Figure 3. Fixation durations (in milliseconds) of read and unread tuples among those whose fixation = 1

4.2 From Reading to Click

According to the examination hypothesis proposed in [5, 27], one result on SERP will be clicked if and only if it is both attractive to and examined by a certain user. This means that examination is a prerequisite for result clicking. Examination may have different kinds of definitions but when users examine results, they should at least read the content of the results. Therefore, we propose the following hypothesis:

Hypothesis 2 (Reading Hypothesis): *Reading a search result is a prerequisite for clicking on the result.*

To verify the *Reading Hypothesis*, we examined the distribution of tuples according to whether they were clicked and whether they were annotated as "read". From the experimental results shown in Table 2 we find that over 98% of the tuples follow the *Reading Hypothesis* (all except those whose click=1 and reading=0). It means that the hypothesis is valid in most cases. The tuples that do not follow the hypothesis can be explained by some users' memory confusions. In any case, the small proportion of such cases shows that the data collected in our experiment is reliable.

Table 2. Distribution of <*user*, *query*, *result*> tuples with respect to reading and clicking behaviors

	Reading=0	Reading=1
Click=0	59.24%	17.57%
Click=1	1.18%	22.01%

We can also find that not all "read" results were clicked by search users since there are 17.57% results whose reading=1 while click=0. Although these results were annotated as "read" (which means users paid much attention to them), users found them not so relevant or attractive through careful reading and decide not to click on them ultimately. This is in agreement with our intuition that not all results which seem to be relevant in the first glance are attractive if we read them more carefully (see Section 6.1).

Table 3 shows the distribution of <*user*, *query*, *result*> tuples according to whether they were fixated and clicked.

Table 3. Distribution of <*user*, *query*, *result*> tuples with respect to fixation and clicking behaviors

	fixation=0	fixation=1
Click=0	34.96%	41.85%
Click=1	2.15%	21.04%

From the proposed *Fixation Hypothesis* and *Reading Hypothesis*, we can conclude that eye fixation is necessary for reading and reading is necessary for clicking. Therefore, it is reasonable to assume that eye fixation is also necessary for click. From Table 3 we can see that the conclusion holds since the majority of the tuples that were clicked were also fixated.

4.3 Two-Stage Examination

Given the proposed two hypotheses, we find that the result examination process should not be regarded as equal to the eye fixation process. On the one hand, fixation is a prerequisite for reading and reading is necessary for clicking; on the other hand, there are a large number of cases where users fixated on several results but did not consider them as "read" (see Table 2). From the skimming to reading process, users try to decide which results should be paid more attention to and which ones deserves no more future attentions. We believe that the two-stage examination process corresponds to the selective attention mechanism described in [33, 34] as a "bottom-up, spatially parallel process of unlimited capacity". It is also closely related to the information triage process described in [37]. However, this process has not been investigated previously in Web search. The proposed model can be viewed as a special case of the selective attention mechanism and the information triage process.

Our analysis shows that the examination process is not a trivial process and a two-stage model (Figure 4) can better fit it.

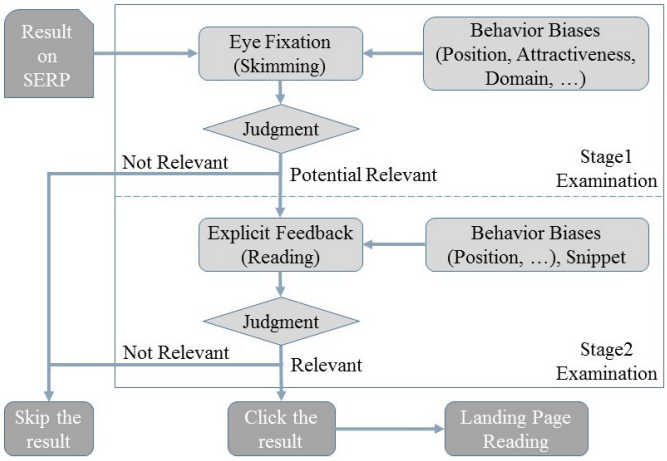

Figure 4. A two-stage examination model which contains a "from skimming to reading" stage (Stage 1) and a "from reading to clicking" stage (Stage 2)

In Figure 4, the examination process is divided into two stages: a "from skimming to reading" stage (Stage 1) and a "from reading to clicking" stage (Stage 2). Each stage involves certain kinds of user judgment on result relevance and the signals adopted in each stage's judgment are different (see Section 5 for details).

According to the two-stage model, the examination process is described as a *skimming-reading-clicking* process. Different from the single-stage *fixation-click* process adopted by most existing works, this proposed model features a reading step, which is captured by the experimental system through user feedback. The necessity of introducing the reading stage is based on the following two findings revealed in our experiments.

First, fixation does not always lead to reading and comprehension of result content according to Section 4.1. Therefore, fixation cannot be simply regarded as examination as in most existing studies. Collecting information on reading as in our experiment will help us better understand the examination process of search users. Since the reading behavior is different from fixation, it is necessary

to collect such information and construct more reliable behavior models for the estimation of examination probability.

Second, the purposes of the two examination stages are not exactly the same (see Section 5 for more details). In Table 2, although users fixated on a relatively large proportion (62.90%) of results, they think that they have only read about half (54.20%) of them. This means that after a user fixates on a certain result, he/she has to decide (not always consciously) whether this result is worth a careful reading or can just be ignored. Users have to make this judgment within a very short time since the median fixation length of results that were fixated but not "read" is about 1.5 seconds according to our experiment. It is known that the reading speed is approximately 200 milliseconds per word [25] in English and 250-350 milliseconds every 2-4 characters [20, 26] in Chinese, the information acquired from Stage 1 is rather limited. Therefore, the judgment in Stage 1 does not heavily rely on comprehension of detailed snippet textual contents. It aims to reduce the result set by discarding those that are obviously irrelevant. Meanwhile, in Stage 2, relevance judgment has to be made by content comprehension and it aims to select the truly attractive ones.

Compared with the traditional single-stage model, our two-stage model seems to fit better the user's actual cognitive process in Web search. It is now important to examine the signals that affect users' examination behavior in the two stages. Such analyses may help develop more reasonable user behavior models.

5. BEHAVIOR BIASES IN TWO STAGES

Previous studies have shown the existence of a number of search behavior biases including position bias [5], domain bias [13] and attractiveness bias [1, 22]. According to the widely adopted examination hypothesis, the fact that one result is clicked after being examined is solely determined by its relevance. Therefore, these behavior biases are regarded as factors affecting users' examination processes. We follow this assumption and focus on how these biases influence users' judgment in the proposed two-stage examination model.

We first provide some definitions that will be used in the following sections. After a user submits a query q to a search engine, he/she will receive a search result page (SERP) which contains a number of search results. While the user examines the SERP, the probability of fixating on a certain result s_i is denoted as $P(F_i)$. The probability of annotating s_i as "read" is $P(R_i)$. After the examination process, the probability of whether the user clicks on s_i is denoted as $P(C_i)$. For the traditional single-stage examination model, user's relevance judgment can be estimated by $P(C_i \mid F_i)$ according to examination hypothesis. Meanwhile, for the proposed two-stage model, the judgment in Stage 1 could be described by the conditional probability $P(R_i \mid F_i)$ and the judgment in Stage 2 is formulated as $P(C_i \mid R_i)$

5.1 Position Bias

The existence of position bias in Web search is validated by a number of existing click-through and eye-tracking studies [5, 10, 18]. It assumes that higher-ranked results receive more user attention and larger probabilities of examination during search sessions. Most click-related studies [4, 5, 8, 35] tried to propose methods to estimate the probability of examination with regard to result position. In the proposed two-stage model, we want to find out how result's ranking position affects the examination behavior and user's judgment in two separate stages. In figure 5, we show how the values of $P(R_i \mid F_i)$ and $P(C_i \mid R_i)$ vary with respect to result position for all relevant results in our data set described in Section 3. We choose the distribution on relevant results instead of all results because the users' relevance judgments are largely affected by the result's actual relevance. It would be unreasonable to compare users' judgment on both relevant and irrelevant results at the same time.

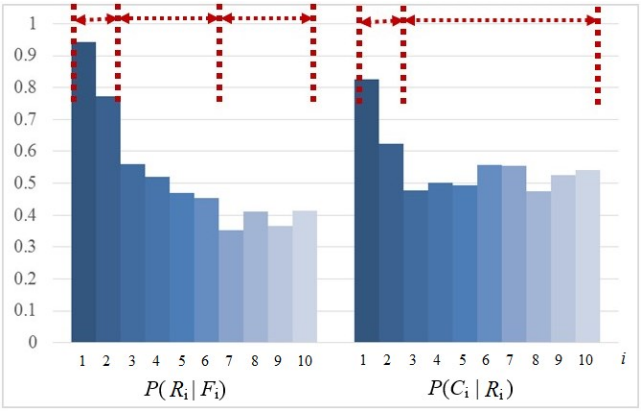

Figure 5. Distribution of $P(R_i \mid F_i)$ and $P(C_i \mid R_i)$ for relevant results with different result rankings.

$P(R_i \mid F_i)$ and $P(C_i \mid R_i)$ can be regarded as users' judgments for result relevance in Stage 1 and Stage 2, respectively. From the figure we can see that the judgment processes in both stages are affected by result ranking positions. Although all results considered here are relevant, users' judgments on different positions are quite different according to Figure 5. The top two results are significantly favored by users in both stages (with t-test p-value<0.001 compared with other results). Users tend to have a larger probability to read and then click on them even when other results are equally relevant.

Figure 5 also shows that the position factor affects two examination stages in different ways. In particular, for the results from 3rd to 10th positions, the examination behavior in Stage 2 is not as strongly biased toward higher ranking positions as in Stage 1. There is no significant difference in $P(C_i \mid R_i)$ between results from 3rd to 6th positions and results from 7th to 10th positions. Meanwhile, for examination behavior in Stage 1, $P(R_i \mid F_i)$ drops approximately linearly from the 3rd to the 6th positions and then remains relatively stable for the rest of the results (the difference is significant with t-test p-value<0.001). We can also find that the differences between the top two results and the other results are not as large for examination behavior in Stage 2 as that in Stage 1 (although both differences are significant).

The above observation confirms that position bias affects users' examination behaviors in both stages. In Stage 1, users quickly skim the results for possible interesting ones and position plays an important role in deciding which results should be retained. In Stage 2, users tend to read the snippets with more care and the position factor seems less important, except for the first two results, which are still more likely to be clicked by users. While our general observation of position bias is consistent with existing researches, we further show that its effects in two stages are different.

5.2 Domain Bias

Domain bias is proposed first by Ieong et al. in [13]. It focuses on the behavior bias of search users on results from different Web domains. The domain bias hypothesis states that the results from trust-worthy Web domains are preferred by users. To validate this hypothesis on our data set, we define "trust-worthy Web domains" as those ranked among the top 100 popular domains according to Alexa China. We choose the ranking of Alexa China instead of its global ranking because most participants in our experiment are

from China and they are more familiar with domains within the Chinese Web.

According to the statistics in Table 4, we can see that this domain bias factor has a major effect on examination behavior in Stage 1: The average $P(R \mid F)$ of results from reputable domains is significantly larger than that of results from other domains. This confirms that users prefer reputable results in their judgment of Stage 1. However, this observation does not hold for the examination behavior in Stage 2: $P(C \mid R)$ values of reputable and other results are almost the same and no statistical significance is observed between them.

Table 4. Comparison of users' examination behavior on reputable results (results from Alexa China's top 100 popular domains) and other results in two stages

		Results from reputable domains	Results from other domains
$P(R \mid F)$	Average	0.6134	0.5194
	Variance	0.0658	0.0799
	p-value	0.0007	
$P(C \mid R)$	Average	0.4708	0.4737
	Variance	0.0637	0.0893
	p-value	0.3119	

The numbers in Table 4 confirm the fact found in previous studies, that the domain bias affects users' examination behavior. Furthermore, we show that this effect only happens to relevance judgments in examination Stage 1 and almost disappears in Stage 2. That is, users tend to trust results from reputable results at their first glances and are more likely to read them carefully. However, the final relevance judgment on whether to click or not is hardly affected by the domains of results.

This finding can be used to adjust result ranking strategies. Although some existing studies such as [13] show that results from reputable domains are preferred by users, this preference should be reconsidered within our two-stage model.

To investigate the relationship between position bias and domain bias in the two-stage examination framework, we show the values of $P(R \mid F)$ for reputable results and other results in different ranking positions in Figure 6.

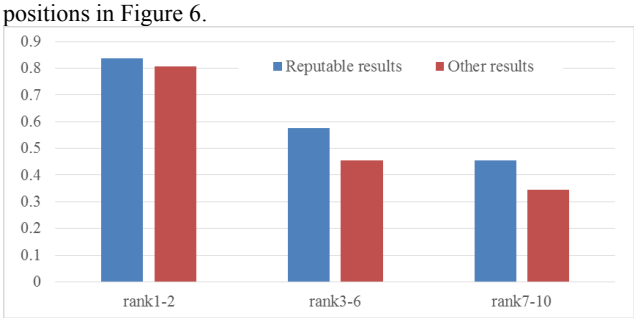

Figure 6. $P(R \mid F)$ values of reputable results and other results at different ranking positions

We separate the results from different ranking positions into three categories (1-2, 3-6, 7-10) and show how $P(R_i \mid F_i)$ at different positions is affected by the domain bias in Figure 6. We can see that results from the top 2 ranking positions are not largely affected by the domain bias factor since the difference between reputable results and other ones is quite small. However, the differences in other ranking positions are significant and results from reputable domains are much more preferred. This phenomenon may be explained by the fact that the top 2 results are favored by users in both stages (see Figure 5). For top 2 results, users tend to judge them without considering which domain they come from. In other words, position bias seems to play a more important role in user examination than domain bias in Stage 1.

5.3 Attractiveness Bias

Attractiveness bias in search has been investigated by a number of researchers [1, 22, 35]. It is found that exact match in result titles and abstracts (which is usually shown in a different color or in bold) affects user judgment. To examine the attractiveness bias in the two-stage examination model, we define the results with the longest exact match in title in a SERP as the attractive ones. In the SERPs of our experiment, the exact matched keywords are shown in a different font color (in red) just as in most commercial search engines. Therefore, the attractive results usually appear with titles almost fully in red. We also tried a number of other definitions of "attractive results" such as those with the longest exact matches in snippets, the proportion of matching terms in snippets/titles, but these definitions did not show significant difference with the current definition.

Table 5. Comparison of users' examination behavior on attractive results (results with longest title exact match) and other results in two stages

		Attractive results	Other results
$P(R \mid F)$	Average	0.6373	0.4846
	Variance	0.0588	0.0660
	p-value	0.0058	
$P(C \mid R)$	Average	0.5778	0.4725
	Variance	0.1226	0.0827
	p-value	0.1585	

From Table 5, we can see that attractiveness bias also plays different roles in users' examination behavior in two stages. The significant difference (two-tailed t-test *p*-value is 0.0058) between average $P(R \mid F)$ values of attractive results and other less attractive ones indicates that attractiveness bias only affects the examination behavior in Stage 1. Although there are also some differences on $P(C \mid R)$ between attractive results and other ones, the differences are not significant. Similar to domain bias shown in Table 4, this observation suggests that the attractiveness in result appearance leads to more user preference in Stage 1, while after more careful reading, attractiveness does not affect users' click behaviors in Stage 2. Once again, we see that the final judgment on whether to click on a result is mainly based on relevance and merely affected by the domain or appearance of the results.

We also examine the relationship between position bias and attractiveness bias in the two-stage examination framework as shown in Figure 7.

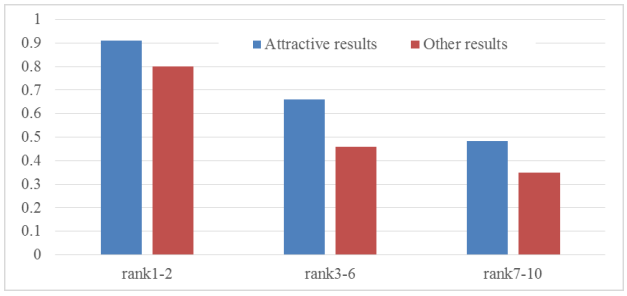

Figure 7. $P(R \mid F)$ values of attractive results and other results in different ranking positions

We can see that results at different ranking positions are all affected by the attractiveness bias, but for top-ranked results the effect is not significant. Together with the observation in Section 5.2, we can

conclude that users tend to judge the top 2 results by their potential relevance no matter which domain they come from and whether they are attractive or not.

5.4 Findings on Behavior Biases

From the observations on behavior biases, we found that different biases affect different examination stages. Position bias plays the most important role because they affect users' relevance judgment in both stages and the top-ranked results seem not to be significantly affected by other biases. Domain bias and attractiveness bias both have significant impact on users' examination behaviors in Stage 1, while in Stage 2, reputable or attractive results are not favored as much as they are in Stage 1.

The fact that behavior biases affect in different ways the two examination stages validates the necessity of constructing a two-stage examination model. We indeed observed that users behave differently in those two stages and this should be captured in a better user examination model. In section 6, we will further demonstrate the necessity of this from the relevance prediction perspective. As stated in previous sections, these findings could help us improve search ranking strategies by incorporating the biases accordingly. For example, although domain bias and attractiveness bias exist in user behaviors, after a more careful reading, users do not favor reputable or attractive results to click on compared to other ones. Consequently, if reputable or attractive results are marginally relevant but placed at high positions, users may have to make more efforts in examining them but may not click on them.

These findings may also help us to design more reasonable metrics for search performance evaluation. Considering the fact that position bias has little effect in users' relevance judgment in Stage 2 for results at 3rd to 10th positions (as shown in Figure 5), the decaying factor of result position in evaluation metrics should be re-designed in accordance with users' actual examination behaviors. The two-stage model also provides us with more insights in the utility of the user's examination efforts for a given ranking. When the relevance label, the site reputation and attractiveness scores are available, we could train a two-stage model to better estimate the expected effort user could spend and the expected utility he will gain. Therefore, the two-stage model can inspire the design of more reasonable evaluation methodologies, which we leave as future work.

6. RELEVANCE ESTIMATION

6.1 Relevance Judgments in Two Stages

As the comprehension degree of the user in each examination stage is different, his/her action also has different implication with respect to relevance. In Stage 1, through quick skimming, the user makes a decision on whether to further examine the result. If the user considers the result as potentially relevant, he/she will continue reading the result and go into Stage 2. While in Stage 2, after a more careful reading of the result snippet, the user makes a decision on whether to click it. Therefore, we could use the manually assessed relevance judgments described in Section 3.3 to evaluate how a selection at each examination stage entails relevance. In Stage 1, we consider the action that subjects label a certain result as "read" as a selection. In Stage 2, we took the action that subjects clicked on a certain result as a selection. Table 6 shows the comparison of relevance implication of the two different stages in the proposed model.

According to the statistics in Table 6, the number of examined results in Stage 2 (as well as recall value) is much smaller than that in Stage 1. This is in agreement with the selective attention mechanism [33] that the "early, pre-attentive stage" has a much higher coverage of cases than the "later, attentive limited-capacity stage". Meanwhile, the evaluation numbers also show that the selection made in Stage 2 is more accurate and reliable than that in Stage 1 with respect to relevance. As stated in Section 4.1, the average length of fixation on tuples in Stage 2 is much larger than that in Stage 1. One may expect a higher implication of relevance judgment by user's action in Stage 2 because users pay more attention to the results.

Table 6. Relevance implication in two examination stages. The signs (+, -) show the comparison of Stage 2 to Stage 1

	Stage 1	Stage 2
#Examined	5,600	3,035 (-45.80%)
#Relevant	3,446	2,111 (-38.74%)
Accuracy	0.5966	0.6415 (+7.53%)
Precision	0.6955	0.7875 (+13.22%)
Recall	0.6126	0.4066 (-33.63%)
KAPPA	0.1773	0.2312 (+30.42%)

From Stage 1 to Stage 2, the number of examined results drops from 5,600 to 3,035 while the percentage of relevant results rises from 61.54% to 69.56%. This result is intuitive: users examine a large number of results for a quick filtering in Stage 1, while more careful selection of relevant results is made in Stage 2.

6.2 Predicting Two-Stage Examination with Mouse Movements

Extracting implicit relevance feedback information from user behavior is one of the goals of search ranking researches. To do this, one has to collect eye-tracking data and users' explicit feedback on results. However, eye-tracking devices are quite expensive and data can only be collected for a relatively small number of subjects (typically tens of subjects). To make the two-stage examination model usable in practical Web search environment, we have to rely on signals available at large scale. Mouse movements are such signals. Mouse movement features can be collected at large scale and they have been used in a number of existing studies to predict eye movement [15], estimate relevance [32] and improve click models [16]. The previous studies showed strong correlation between eye-tracking data and mousse movements. This provides evidence that our two-stage examination model can be adapted to mousse movement data. With the experimental system described in Section 3, we collected mouse movement information for all tuples with injected JavaScript on SERPs. A number of mouse movement features, as shown in Table 7, are extracted and adopted to predict the examination behavior in two stages.

Table 7. Mouse movement features adopted for predicting two-stage examination behaviors

Feature	Description
Distance	The total cursor movement distance
MovePosition	The leftmost/rightmost/upmost/bottommost position cursor ever reaches
MoveDistance	The total leftwards/rightwards/upwards/ downwards movement distance
Scroll	Whether user scrolls up / down to the result
MouseTime	Total mouse dwell time on the result
SearchTime	Time user spends on the corresponding SERP
TotalTime	Time user spends on the whole search task
ArrivalTime	Time elapsed until the result is hovered for the first time

As shown in Table 7, some of the behavior features (e.g. MouseTime, ScrollUp, ScrollDown) have been validated by existing studies [32] as important signs for user examination. We

also include a number of new features, which correlate with eye fixation or reading feedback according to our data set.

A number of learning-based classifiers are trained to estimate the examination probabilities in both stages for each tuple. In the classification process, we group the tuples in the dataset described in Section 3 into the following three categories (note that tuples which were fixated for less than 500 milliseconds but annotated as "read" are removed because they are regarded as noise):

- ***Not examined*** **(E0)**: tuples which were fixated for less than 500 milliseconds (following the threshold setting in Section 4.1)
- ***Examined in Stage 1*** **(E1)**: tuples which were fixated for no less than 500 milliseconds while not labeled as "read".
- ***Examined in Stage2*** **(E2)**: tuples labeled as "read".

With different learning algorithms, we predict the label of each tuple as E0, E1 or E2. The prediction performance of these classifiers is compared based on a five-fold cross validation on the dataset. We make sure that the training and test sets do not share any query in common. Based on the comparison results, the Gradient Boosted Regression Trees (GBRT) method is found to perform the best on most metrics (accuracy is ***0.6393*** and Kappa coefficient is ***0.4519***). The prediction of examination behavior with the proposed mouse movement features achieves relatively high performance (Kappa value shows substantial agreement), which means that the two-stage examination behaviors could be identified in practical Web search environment with mouse movement information. This result is consistent with findings in previous studies that gaze and mouse movement behaviors are highly correlated [15, 28].

6.3 Result Relevance Estimation

To extract relevance feedback information from both users' click-through and mouse movement logs, we follow the examination hypothesis and treated $P(R \mid F)$, $P(C \mid R)$ and $P(C \mid F)$ as signals of relevance judgment. Among these signals, $P(R \mid F)$ and $P(C \mid R)$ are new features extracted based on the two-stage model while $P(C|F)$ is extracted based on the original single-stage examination model. To show the effectiveness of the proposed model in practical Web search environment, we adopted both the actual and predicted user behavior on fixation (F) and reading (R) in the calculation of these signals. GBRT is employed to generate the prediction of user examination behavior since it gains best performance.

Table 8 shows the relevance estimation performance of both the proposed two-stage examination model and the original single-stage model. Relevance estimation with both actual user behavior and predicted behavior are also compared. As for the two-stage model, $P(R \mid F)$ and $P(C \mid R)$ are both used to estimate result relevance while for single-stage model, $P(C \mid F)$ is used as the sign for relevance of results. Notice that in Table 8, $P(R \mid F)$, $P(C \mid R)$ and $P(C \mid F)$ are all predicted with the same classifier and the mouse behavior features used in both the predicted two-stage model and the single-stage model are the same. As for learning method adopted in relevance estimation, we compare several different algorithms and choose SVM to combine the extracted features.

Results in Table 8 show that relevance estimation based on the proposed two-stage examination model outperforms single-stage model significantly in terms of Accuracy, F-measure and Kappa coefficient (t-test p-value<0.001). We can also find that relevance estimation results with the actual user behavior are slightly better than those with the predicted behavior, probably because there are possible errors in the predicted results. Although the prediction of examination behavior do not reach the level of perfect agreement according to Kappa and accuracy values, the differences between relevance estimation results of actual and predicted behavior are not significant. This means that relevance estimation with mouse movement and the two-stage examination model can achieve comparable results with the estimation based on eye-tracking and user feedback information. We cannot conclude that the expensive eye-tracking information could be replaced by mouse movement data because only eye fixation length on certain results are used in our work instead of detailed eye movement behavior data. However, it does show that mouse movement data which could be collected at large scale is good enough for predicting the examination behavior proposed in our two-stage model and improves the estimation of result relevance.

Table 8. Relevance estimation performance of two-stage and single-stage models with actual/predicted user behaviors

	Actual User Behavior (incl. eye movement, user feedback on reading)		Predicted Behavior (mouse movement information only)	
	Two-stage model	Single-stage model	Two-stage model	Single-stage model
Accuracy	0.6440	0.5760	0.6400	0.5720
Precision	0.6910	0.8221	0.6872	0.8155
Recall	0.6970	0.3356	0.6941	0.3345
F-measure	0.6865	0.4747	0.6799	0.4693
Kappa	0.2727	0.2141	0.2688	0.2052

In summary, different from most existing studies which consider fixation behavior as examination, the proposed two-stage model introduces reading behavior of search users. From the experimental results in Table 6, the relevance judgment made in Stage 2 is more accurate while that in Stage 1 covers more search results. Although relevance estimation based on predicted two-stage examination behaviors does not involve extra information besides mouse click-through and movement behavior, it outperforms the estimation by predicted single-stage behavior. This result could be explained by the fact that with the single-stage model, all fixated but not clicked results will be regarded as "irrelevant", while in the two-stage model, the classifier will make the judgment according to whether user has read and comprehended it. If the user has not comprehended the results, we cannot simply judge it as irrelevant because the judgment in Stage 1 may not be reliable. All the experimental results indicate that the proposed model could better describe users' cognitive behavior in search environment and better extract relevance feedback information from users' behavior.

7. CONCLUSIONS AND FUTURE WORK

In this paper we investigated user's examination behavior in Web search. By conducting carefully designed experiments we found that user's examination process could be separated into two different stages. Similar to the attention selection mechanism and information triage process described in many cognitive studies, in Stage 1 users quickly skim a relatively large number of results for possible interesting ones; while in Stage 2 they read a limited number of result snippets more carefully and make click decisions. Experimental analyses show that these two examination stages have different behavior biases as well as relevance judgment implications. As mouse movement information could be collected at large scale in practice, we also tried to use mouse movements for the prediction of the two-stage examination and the estimation of result relevance. Results show that we can achieve a better relevance estimation with the proposed two-stage model than the original single-stage model. These findings show that the proposed model can better reflect the user's cognitive behavior and present a new way to combine mouse movement information into relevance

estimation. Several further aspects are interesting to explore in the future. For example, one can construct click models based on the two-stage examination framework to improve search ranking performance. We can also test the model in real Web search environment by involving a larger number of participants and by consodering vertical/sponsored search results.

8. ACKNOWLEDGEMENTS

Part of the work has been done at the Tsinghua-NUS NExT Search Centre, which is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490). This work is supported by Tsinghua-Samsung Joint Lab, National Key Basic Research Program (2015CB358700) and Natural Science Foundation (61472206, 61073071) of China.

9. REFERENCES

[1] J. Bar-Ilan, K. Keenoy, M. Levene, and E. Yaari. Presentation bias is significant in determining user preference for search results: a user study. JASIST, 60(1):135–149, 2009.

[2] A. Broder. A taxonomy of web search. ACM SIGIR forum, 36: 3–10. ACM, 2002.

[3] G. Buscher, R. W. White, S. Dumais, and J. Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In WSDM, pp. 373–382. 2012.

[4] O. Chapelle and Y. Zhang. A dynamic Bayesian network click model for web search ranking. In WWW, pp. 1–10. 2009.

[5] N. Craswell, O. Zoeter, M. Taylor, B. Ramsey. An experimental comparison of click position-bias models. In WSDM, pp. 87–94. 2008.

[6] E. Cutrell and Z. Guan. What are you looking for: an eye-tracking study of information usage in web search. In CHI, pp. 407–416. ACM, 2007.

[7] F. Diaz, R. White, G. Buscher, and D. Liebling. Robust models of mouse movement on dynamic web search results pages. In CIKM, pp. 1451–1460. ACM, 2013.

[8] G. E. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In SIGIR, pp. 331–338. ACM, 2008.

[9] J. H. Goldberg and J. I. Helfman. Comparing information graphics: a critical look at eye tracking. In Proceedings of the 3rd BELIV'10 Workshop. 71–78. ACM, 2010.

[10] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. SIGIR2004. 478–479.

[11] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In CHI, pp. 417–420. 2007.

[12] Q. Guo and E. Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In WWW 2012. pp.569–578. 2012.

[13] S. Ieong, N. Mishra, E. Sadikov, and L. Zhang. Domain Bias in Web Search. In WSDM, 2012.

[14] A. J. Hornof. Cognitive strategies for the visual search of hierarchical computer displays. Human-computer Interaction, 19(3):183–223, 2004.

[15] J. Huang, R. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. CHI2012. 1341–1350.

[16] J. Huang, R. White, G. Buscher, K. Wang. Improving searcher models using mouse cursor activity. SIGIR'12. 195–204.

[17] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: using cursor movements to understand and improve search. In CHI, pp. 1225–1234. 2011.

[18] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting click-through data as implicit feedback. In SIGIR, pp. 154–161. ACM, 2005.

[19] M. A. Just, P. A. Carpenter. A theory of reading: From eye fixations to comprehension. Psychological review, 87:329, 1980.

[20] X. Li, P. Liu, and K. Rayner. Eye movement guidance in Chinese reading: Is there a preferred viewing location? Vision Research, 51(10):1146–1156, 2011.

[21] L. Lorigo, M. Haridasan, H. Brynjarsdóttir. Eye tracking and online search: Lessons learned and challenges ahead. JASIST, 59(7):1041–1052, 2008.

[22] Y. Yue, R. Patel, and H. Roehrig. Beyond position bias: examining result attractiveness as a source of presentation bias in click-through data. In WWW, pp. 1011–1018. 2010.

[23] V. Navalpakkam, L. Jentzsch, R. Sayres. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In WWW, pp. 953–964, 2013.

[24] M. Yan, R. Kliegl, E. M. Richter, A. Nuthmann, and H. Shu. Flexible saccade-target selection in Chinese reading. The Quarterly Journal of Experimental Psychology, 63(4):705–725, 2010.

[25] K. Rayner, A. Pollatsek, J. Ashby, and C. Clifton. Psychology of reading (2nd Edition). 2011.

[26] E. M. Reingold, E. D. Reichle. Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. Cognitive psychology, 65(2):177–206, 2012.

[27] M. Richardson, E. Dominowska, R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In WWW. pp. 521-530, 2007.

[28] K. Rodden, X. Fu, and A. Aula. Eye-mouse coordination patterns on web search results pages. CHI2008. 2997–3002.

[29] D. D. Salvucci, J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In Symposium on Eye tracking research & applications, pp. 71–78. 2000.

[30] R. Shiffrin, W. Schneider. Controlled and automatic human information processing: perceptual learning, automatic attending and a general theory. Psychological review, 84:127, 1977.

[31] N. C. Smeeton. Early history of the kappa statistic, 1985.

[32] M. Speicher, A. Both, and M. Gaedke. Tellmyrelevance!: predicting the relevance of web search results from cursor interactions. In CIKM, pp. 1281–1290. 2013.

[33] J. Theeuwes. Visual selective attention: A theoretical analysis. Acta Psychologica, 83:93–154, 1993.

[34] J. Theeuwes, A. F. Kramer, S. Hahn, and D. E. Irwin. Our eyes do not always go where we want them to go: Capture of the eyes by new objects. Psychological Science, 9(5):379–385, 1998.

[35] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, K. Zhang, Incorporating vertical results into search click models. In SIGIR 2013.

[37] E. Niebur, L. Itti, and E. Koch. Controlling the focus of visual selective attention. In Models of neural networks IV, pp. 247–276. New York: Springer. 2001.

[38] M.D. Smucker, X.S. Guo, and A. Toulis. Mouse movement during relevance judging: implications for determining user attention. In SIGIR2014. 979-982.

[39] Cole M., Hendahewa C., Belkin N., & Shah C. Discrimination Between Tasks with User Activity Patterns During Information Search. In SIGIR 2014.