Advances in Computational Intelligence (2022) 2:14
https://doi.org/10.1007/s43674-021-00028-2

ORIGINAL ARTICLE

From linear to non-linear: investigating the effects of right-rail results on complex SERPs

Yunqiu Shao[1] · Jiaxin Mao[2] · Yiqun Liu[1] · Min Zhang[1] · Shaoping Ma[1]

Received: 22 June 2021 / Revised: 11 October 2021 / Accepted: 24 November 2021

Abstract

Modern search engine result pages (SERPs) become increasingly complex with heterogeneous information aggregated from various sources. In many cases, these SERPs also display results in the right rail besides the traditional left-rail result lists, which change the linear result list to a non-linear panel and might influence user search behavior patterns. While user behavior on the traditional ranked result list has been well studied in existing works, it still lacks a thorough investigation of the effects caused by the right-rail results, especially on complex SERPs. To shed light on this research question, we conducted a user study, which collected participants' eye movements, detailed interaction behavioral logs, and feedback information. Based on the collected data, we analyze the influence of right-rail results on users' examination patterns, search behavior, perceived workload, and satisfaction. We further construct a user model to predict users' examination behavior on non-linear SERPs. Our work contributes to understanding the effects of the right-rail results on users' interaction patterns, benefiting other related research, such as the evaluation and UI optimization of search systems.

Keywords Web search · Search result page · Eye-tracking · User modeling

1 Introduction

Modern search engine result pages (SERPs) are far more complex, composite, and informative than traditional ten blue links. Along with algorithmic web results (called *organic* results), heterogeneous *vertical* results, such as images, videos, news, answer cards, are aggregated on the SERPs. Beyond a *linear* result list, the SERP has evolved into a *non-linear* layout. As illustrated in Fig. 1, the layout of a modern SERP can generally be divided into two parts, a left rail containing both organic and vertical results in the form of a ranked list, and a right rail, which usually displays the results given by semantic retrieval techniques Bota et al. (2016). Different search engines put different results in the right rail. Figure 1 illustrates two typical right-rail layouts, which are both widely adopted by commercial search engines. One is a composite entity card that aggregates elements extracted from various resources, such as images, textual information, and entities. Figure 1a is an example of the composite entity card from *Google*. The other layout presents results in a blocked-based way. Taking Fig. 1b (from *Sogou*) as an example, different categories of related entities are divided into blocks, and these blocks are arranged in a listwise way. Within each block, various meta-information are included, e.g., image and brief description. Besides, an entity graph is sometimes involved in this layout.

Modeling user behavior is fundamental in information retrieval (IR) research and beneficial to improving many IR-related tasks Agichtein et al. (2006), Bron et al. (2013), Zhang et al. (2020), Zhang et al. (2021). Although a large number of existing works have studied user behavior in the desktop environment from different aspects Rele and Duchowski (2005), Hotchkiss et al. (2005), Zhang et al. (2020), Sakai and Zeng (2020), most of them are focused on a single ranked list of search results. Notably, previous work Hotchkiss et al. (2005)

✉ Yiqun Liu
yiqunliu@tsinghua.edu.cn

Yunqiu Shao
shaoyunqiu14@gmail.com

Jiaxin Mao
maojiaxin@gmail.com

Min Zhang
z-m@tsinghua.edu.cn

Shaoping Ma
msp@tsinghua.edu.cn

[1] BNRist, DCST, Tsinghua University, Beijing, China

[2] GSAI, Renmin University of China, Beijing, China

Published online: 10 January 2022

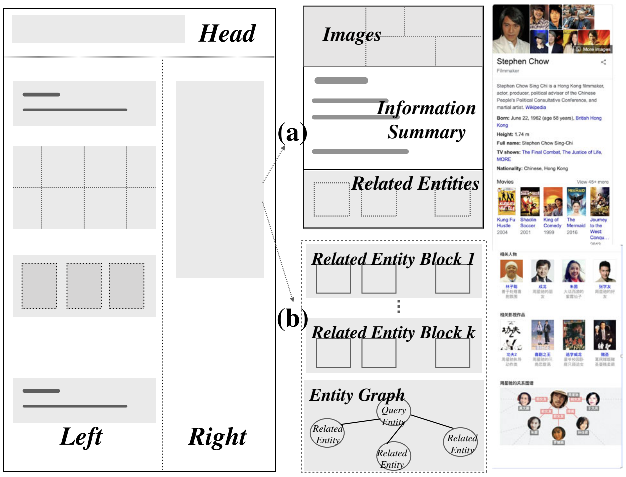


Fig. 1 Illustration of the modern non-linear SERP. **a** and **b** Are two typical layouts of the right rail. **a** Is a composite entity card. **b** Is a block-based layout, and this example contains both entity blocks and an entity graph

observed a "Golden Triangle" pattern of examination attention in the linear layout composed of organic results, where users pay more attention to the top-left results and examine results in a top-down manner. This behavior pattern also leads to the popularly adopted cascade model, within which, related research (e.g., result ranking, search evaluation) have placed extra emphasis on the top-left results.

The evolution of the SERP layout inevitably changes user behavior and, therefore, introduces new challenges for the existing methods. With the development of aggregated web search, a variety of vertical results have been presented in the SERPs. The earlier research line of the complex SERP is focused on the vertical results in the left rail, including user behavior modeling Liu et al. (2015), ranking algorithms Chen et al. (2012), Wang et al. (2013), and evaluation Zhou et al. (2012).

Beyond changes in the left rail, results are also aggregated in the right rail, which changes the SERP into a non-linear layout. In recent years, a few works have paid attention to the non-linear layout. Specifically, Navalpakkam et al. (2013) pointed out that the flow of user attention on the non-linear page would change with the existence of the knowledge graph (KG) on the right side. Bota et al. (2016) investigated the effects of the composite entity cards through a crowdsourced user study. However, prior studies are focused on the composite entity card (Fig. 1a) in the right rail, ignoring the block-based right rail, which is also popularly adopted in the current search systems. Another limitation lies in that they mostly suppose that the results in the left rail are all organic ones, but a large proportion of these results are vertical ones in practical settings. Many previous studies have shown that the interaction patterns on vertical results are somewhat different from those on organic ones Sushmita et al. (2010), Liu et al. (2015), Arguello and Capra (2014), Wu et al. (2020). Therefore, it is also reasonable to assume that the influence of right-rail results in practical search systems might differ from prior studies.

In the recent research on the complex SERP, a large proportion of works Wu et al. (2020), Zhang et al. (2020), Zhang et al. (2021), Sakai and Zeng (2020) still focus on the left rail, omitting the right rail. For example, Wu et al. worked on the influence of the answer card shown on the top of the left rail Wu et al. (2020). When constructing user models for evaluating the complex SERPs, the right-rail results are filtered out or ignored Zhang et al. (2020), Sakai and Zeng (2020). Notably, Azzopardi et al. (2018) inferred an examination sequence on the non-linear SERP according to the click-through logs of a commercial search engine and developed an evaluation metric based on the Information Foraging Theory (IFT) recently. Based on the examination sequence, Thomas et al. (2018) developed "card-aware" metrics for offline evaluation of card-based SERPs. Their research has also highlighted the importance of considering the effects of right-rail results. However, the examination sequence inferred from click-through logs Azzopardi et al. (2018) might still vary from users' actual examination flow. Therefore, we believe the effects of right-rail results on the complex SERPs are still under investigation.

In this study, we focus on investigating the influences of the block-based right-rail results on the complex SERP. Specifically, this study attempts to address the following research questions (**RQs**):

- **RQ1**: How do users allocate examination attention to the non-linear SERPs? When do they examine the right rail?
- **RQ2**: How do the right-rail results affect users' search process, search satisfaction, and perceived workload?
- **RQ3**: Can we predict the examination behavior on the complex SERPs?

To address these research questions, we conducted a laboratory user study ($N = 30$) using an eye-tracking device, which was carefully designed to simulate a realistic search environment. Eye movement and other interactive signals (e.g., click, hover, dwell time, query formulation) were collected during the search process. Besides, we also collected explicit feedback on search satisfaction and workload from the participants. We mainly investigated the effects of two variables, i.e., the existence and the examination of right-rail results. Specifically, the former one was controlled by manipulating the experimental interface during the search. The latter one was measured with the eye-tracker. The statistical analysis indicates that the display of blocked-based right-rail results has little influence on the search process, perceived satisfaction, or workload. However, users have more interactions with the SERPs, appear to struggle more, and feel less satisfied if they examine the right-rail results. Furthermore, we build an explainable user model to predict examination behavior and make further interpretations. In summary, our key contributions are three folds:

- We conducted an eye-tracking user study tailored for investigating user behavior on the non-linear SERP. Different from previous work Navalpakkam et al. (2013); Bota et al. (2016), this study simulated the realistic web search scenario, involving complex SERPs. This data set does not only support investigating the effects of right-rail results but can also support other IR research related to user modeling on the complex SERP. The data set is now open to the public.
- We investigate the effects of block-based results in the right rail from various perspectives. Compared with click-through, fixations are better indicators of examination, especially for the aggregated results. Based on eye movements and behavioral logs, we characterize how users allocate their examination attention to the complex non-linear SERP. Furthermore, we investigate the influence of displaying and examining the right-rail results on the search process, satisfaction, and workload.
- Inspired by the above observations that users' search behavior and satisfaction will change when they examine the right rail, we propose a supervised learning framework to predict user examination on the right rail using behavioral and static features. Furthermore, we interpret the predictive models by feature analysis, which supports understanding why the user examines the right rail. The results can further benefit other related IR tasks, such as interface optimization and search result evaluation.

2 Related work

2.1 User modeling in web search

User modeling is an essential task in both academic and industrial IR research. User behavior data have been explored and successfully applied to improve a variety of IR tasks in web search, including result ranking Agichtein et al. (2006), Zhang et al. (2021), search evaluation Rele and Duchowski (2005), Zhang et al. (2020), interface optimization Bron et al. (2013), Wu et al. (2020), etc. In the work of user modeling in web search, one of the conventional techniques is to examine large-scale log data of commercial search engines Silverstein et al. (1999), Mat-Hassan and Levene (2005). Besides, interactive data (e.g., mouse clicks and movements) are collected and explored in laboratory user study Huang et al. (2011), Liu et al. (2019) or diary study Teevan et al. (2004), Zhang et al. (2020), Chen et al. (2021). Eye-tracking, which can capture real-time eye movements effectively, is a favored technique for investigating user examination behavior and, therefore, has been utilized in various search scenarios Hotchkiss et al. (2005), Lagun et al. (2014), Xie et al. (2017), Li et al. (2018), Zheng et al. (2020). The "Golden Triangle" pattern was observed on the traditional ten-blue-link pages in web search Hotchkiss et al. (2005). It suggested that users paid most of their attention to the top-left results, and the attention decreased when moving towards the bottom or right of the result page. On the other hand, the "middle-position bias" was observed in image search, where the image results were displayed in a two-dimensional panel Xie et al. (2017), Xie et al. (2019). Besides the layout of the result page, result and task types affect examination patterns as well. Previous studies revealed that users paid more attention to the middle part of the SERP consisting of organic results, ads, and related searches Dumais et al. (2010) and tended to examine the result list deeper and more quickly when facing complex tasks Thomas et al. (2013).

Besides user behaviors during the search process, search outcomes and workload are also significant components of user modeling. For instance, Arguello and Choi (2019) investigated how search outcomes and workload were impacted by individual cognitive characteristics and search interfaces (i.e., interleaved or blocked). Zhang et al. (2020), Zhang et al. (2020) worked on modeling user satisfaction during search.

2.2 Aggregated web search

Modern search engines aggregate heterogeneous results from multiple sources on the result page, known as verticals. Differences in user behaviors have been observed in the context of the aggregated search. Based on the study of user click-through data, Sushmita et al. (2010) found that users clicked more on video results than the image and news ones. Furthermore, Liu et al. (2015) observed attraction, cutoff, and spill-over examination effects brought by vertical results through an eye-tracking user study. Verticals have also been incorporated into click models Chen et al. (2012), Wang et al. (2013) to improve their effectiveness. Recently, Wu et al. (2020) conducted a user study to inspect the impact of providing direct answers in search results, which could also be viewed as a vertical type. In general, most of the research on aggregated search focused on the linear result list shown on the left side of the page.

In recent years, the non-linear layout has also inspired some work on methodologies for interface optimization Wang et al. (2013) and search evaluation Chuklin et al. (2016), Azzopardi et al. (2018), Thomas et al. (2018). Navalpakkam et al. (2013) first observed that the flow of user attention on the non-linear result page with knowledge graphs were different from that on linear pages. However, their study focused more on how to predict the eye gaze on SERPs with mouse movement data and did not thoroughly analyze how users' search behavior was altered by the knowledge graph in the right rail. Bota et al. (2016) conducted a crowd-sourced online user study to investigate the effects of entity cards given ambiguous search topics. However, they only maintained the organic results in the left rail and manipulated the presence and properties of entity cards, including their content, coherence, and diversity. The experimental settings were quite distinct from the realistic search scenario. Recently, Azzopardi et al. fitted an examination sequence on the complex non-linear SERPs, which has been applied for developing evaluation metrics of SERPs Azzopardi et al. (2018), Thomas et al. (2018). However, as stated by Navalpakkam et al. (2013), although mouse measures are somehow correlated with eye gaze, eye-tracking is still more sensitive and cannot be fully substituted. Lacking an in-depth investigation into the impact of right-rail results on the complex SERPs, recent related research mainly dismissed the right rail, such as developing click models Zhang et al. (2021) and evaluating SERPs Zhang et al. (2020), Sakai and Zeng (2020).

This paper conducted a lab-based, eye-tracking study rather than a crowd-sourced study to collect a more comprehensive behavior data set in a more realistic search scenario. In particular, compared with Bota et al. (2016)'s study, we (1) use a SERP layout that is more similar to that of commercial search engines as we have not filtered out the vertical results in the left rail. (2) We cover various types of search tasks instead of limiting them to ambiguous search topics. (3) We not only investigated the influence of displaying right-rail results but also that of examining right-rail results as the eye-tracking data enable us to determine whether the participant has examined the right-rail results. We believe that our study would advance in understanding how users interact with the non-linear SERPs.

3 User study

3.1 Collecting behavior data

Tasks We designed our tasks based on the queries with intermediate frequency in 1 day's search logs of a commercial search engine. We manually selected 21 queries and ensured that all of the initial queries would trigger right-rail results in our experimental search engine. To clarify the information need, we further constructed more detailed background descriptions. The tasks were designed to be either exploring or fact-finding, covering different task types. Table 1 gives examples of each type. Compared with the fact-finding tasks, the exploring tasks usually involved richer information need. We expected users to interact more with the search system, thus including more exploring tasks to collect more behavioral signals in our user study. Specifically, seven of the tasks

Table 1 Examples of search tasks and initial queries for different task types

Task Type	Task Description	Initial Query
Exploring	Imagine that you are planning to travel in Hong Kong for 3 days. You would like to know some places of interest and the possible money cost.	Travel in Hong Kong
Fact-Finding	You would like to download an app called "Himalayan" for the Android phone.	Himalayan

were "fact-finding", and the others were "exploring", and one of the exploring tasks was used as the warm-up training task. Note that we focus on the effects of right-rail results rather than task types in this study. Different task types are involved for better coverage and approximation of the realistic search environment.

Participants We recruited 30 participants (13 males and 17 females, aged from 17 to 33) via online forums and social networks. The participants had a variety of background majors, e.g., engineering, social science, arts. All participants reported being familiar with current search engines in the desktop environment and using web search daily. It took about 1.5 h for each participant to complete the main tasks. The participant would be paid about 15 dollars for involvement.

Experiment system and platform In our study, the experiment system was deployed on a 17-inch LCD monitor with a resolution of 1366 × 768 in pixels. Google Chrome browser was used to display the pages of the experimental systems. We developed a customized browser extension to control the presentation of right-rail results and, meanwhile, log users' search behaviors on the SERPs. The behavioral signals and corresponding SERPs would be saved in a backend database for further analysis. A head-free eye tracker, Tobii X2-30, was applied to capture eye movements. The error angle was no more than 0.5° in our experimental setting, where operating distance was 40–50cm and the gaze angle $\leq 30°$. The eye movements, including fixations and saccades, were detected using the built-in algorithms in Tobii Studio. Specifically, the fixation refers to that the eyes land on an object for a period of time (typically 200–250ms). Meanwhile, the saccade means rapid eye movement from one point to another within a period (typically 20–50ms). Previous research Xie et al. (2017), Li et al. (2018), Zheng et al. (2020) has found that fixations were correlated with user examination attention and involved less noise compared with saccades. Therefore, we mainly inspected the fixations in the following experiments and selected the 200ms as the lower bound of the fixation time, following previous work.

Experimental manipulation To investigate the effects of right-rail results, we manipulated the presence of the right rail in the user study. As shown in Fig. 2, there were two experimental conditions, SERPs with right-rail results (denoted as "w/ R") and SERPs without right-rail results (denoted as "w/o R"). The results shown in the left rail remained identical for both conditions. The injected JavaScript controlled the presence of the right rail according to the experimental condition. Among the 20 main tasks, each participant completed ten under the "w/ R" condition and the other ten under the "w/o R" condition. The main tasks were shown to participants in random order to balance the order bias Lagun et al. (2014). In particular, Fig. 3 gives an example of the "w/ R" SERP. The left rail contained a variety of search results, e.g., organic results and image results. The right rail consisted of blocked-based results, such as related entity blocks and the entity graph. Note that some right rails might not involve the entity graph.

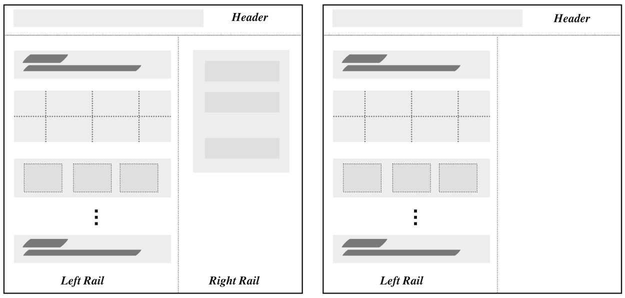


Fig. 2 Illustration of the experimental manipulations, denoted as "w/ R" and "w/o R," respectively

Procedure Figure 3 shows the procedure of the user study. Before starting the main tasks, each participant was instructed to calibrate the eye-tracking device and sign the informed consent. To ensure the participant familiar with the experimental procedure, we used an example task as a tutorial for the warm-up training. After the pre-experiment training, each participant was instructed to complete the 20 main tasks. Each task comprised five stages:

- *Task description* A detailed task description was provided at the beginning of each task to simulate a realistic web search scenario. In this stage, the participant was required to read the description. Then she was asked to repeat it in her own words to make sure having understood the information need.
- *Pre-task questionnaire* Then the participant was asked to report the level of her interest, knowledge, and expected difficulty before searching in the pre-task questionnaire. All the questions were answered in a five-point Likert-type scale (1: not at all, 5: very), following previous work Mao et al. (2018).
- *Search* Once finishing the pre-task questionnaire, the participant was directed to an experimental search system, where the results were crawled from a commercial search engine. In particular, we provided the initial query and the corresponding SERP was crawled and saved in advance. In that way, the first query results were identical for all participants except for the existence of the right rail. In this stage, the participant could freely scroll up and down, click on results, and reformulate queries, just like using a search engine naturally. The participant could finish searching by closing the searching page whenever she thinks that she has obtained adequate information

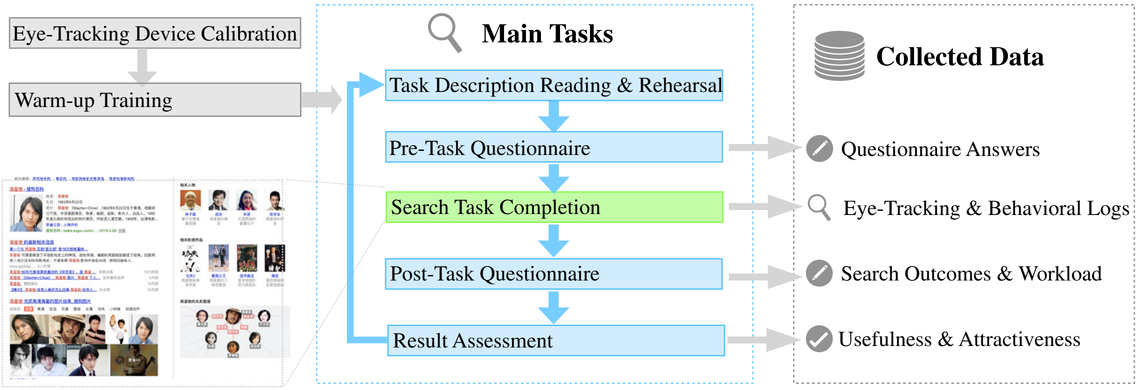

Eye-Tracking Device Calibration
Warm-up Training
Main Tasks
Task Description Reading & Rehearsal
Pre-Task Questionnaire
Search Task Completion
Post-Task Questionnaire
Result Assessment
Collected Data
Questionnaire Answers
Eye-Tracking & Behavioral Logs
Search Outcomes & Workload
Usefulness & Attractiveness

Fig. 3 Procedure of our user study. An example of the non-linear SERP is shown in the lower left corner (some right-rail results only contain entity blocks)

or could hardly find more helpful information for this task. In this stage, the participant's eye movements were recorded by the eye tracker, and her behavioral signals on the SERPs were collected by the chrome extension and the experimental platform.

- *Post-task questionnaire* After the participant completed searching, she was directed to the post-task questionnaire. At first, she was required to answer an open-ended task-specific question to ensure that she did the experiment carefully. Taking the first task in Table 1 ("Travel in Hong Kong") as an example, she was asked to list several potential destinations and give the budget. Then she was instructed to report her perceived workloads Hart and Staveland (1988), the satisfaction of the session and each query. Similarly, these questions were answered in the 5-point Likert-type scales (1: not at all, 5: very).
- *Result assessment* In this stage, the platform showed the corresponding SERPs chronologically. The participant was asked to make annotations of usefulness and attractiveness (both in 4-point scales Mao et al. (2016), Shao et al. (2019)) for the results she has ever examined or noticed in the "search" stage. Once finishing this stage, the participant could start a new task with the same procedure.

Before the experiment, a pilot study, which involved two additional users, was conducted in advance to make sure the procedure and the experimental systems worked well.

Data cleansing After careful inspection, we filtered out the 33 search sessions that might involve technical flaws. To be specific, 3 of them involved incomplete SERPs due to unexpected network instability. 13 of them were caused by the sudden disconnection of the eye-tracking device during the search process. 17 of them were caused by some participants' misoperation which might lead to flaws in recording eye-movement and behavioral logs, including changing their sitting positions drastically (4), searching without running the browser extension plugin (3), ending search without closing the searching pages (10). Table 2 shows the general statistics of the filtered data set. Since about 70% sessions (394 in 567, indeed) contain only one query, we think that the reformulated queries will not have much impact on the behavior on the initial SERP. Following previous works Bota et al. (2016), Navalpakkam et al. (2013) in non-linear SERP, we only consider the initial query of each session in the following query-level analysis for consistency in the result sets across users.

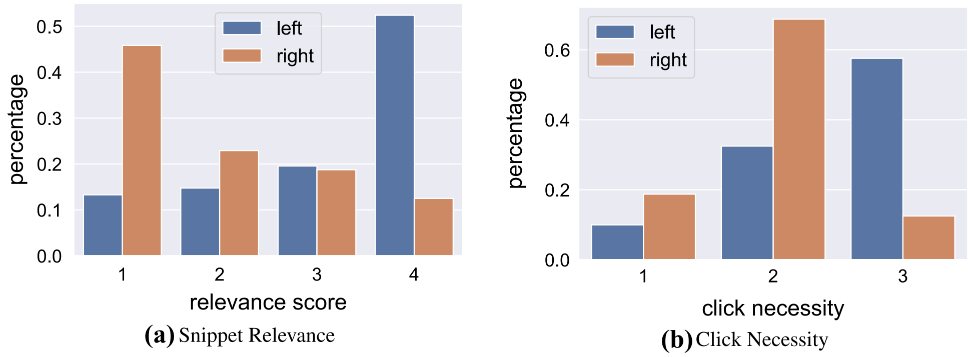
Table 2 Statistics of the collected data set

# users	# tasks	# sessions (w/ R)	# sessions (w/o R)
30	20	281 (E:182, F:99)	286 (E:186, F:100)

"E" and "F" indicate the numbers of "exploring" and "fact-finding" sessions, respectively

3.2 Collecting external annotations

After collecting user behavior and explicit feedback in the user study, we further hired 3 external assessors to make annotations for the results on the first result pages of the initial queries. The results in the left and right rail were annotated in a blockwise way. Task descriptions were also provided for disambiguation. Annotations for "snippet relevance" and "click necessity" Luo et al. (2017) were collected. In particular, snippet relevance of a result was annotated based on its part displayed on the SERP using a 4-point scale Mao et al. (2016), Shao et al. (2019). Click necessity within a 3-point scale (1: not necessary, 2: possibly necessary, 3: necessary) was used to measure the need to click further and visit the landing page, independent of the snippet relevance, following previous work Luo et al. (2017). The Fleiss's Kappa of snippet relevance and click necessity judgments across three assessors are 0.5374 and 0.6107, respectively, indicating moderate agreements Fleiss (1971). If there are disagree-



(a) Snippet Relevance **(b)** Click Necessity

Fig. 4 **a** Distribution of snippet relevance in the left and right rail. **b** Distribution of click necessity distribution in the left and right rail

ments among assessors, the median value is used as the final score in the following experiments.

The distributions of relevance and click necessity are shown in Fig. 4. As for snippet relevance (Fig. 4a), significantly different distributions can be observed in the left and right rail. Over 50% of left results are of high snippet relevance, while a large proportion of the right-rail results are judged to be less relevant. Meanwhile, right-rail results show a lower level of click necessity (Fig. 4b), which suggests that it is easier for users to obtain information or make judgments in the right rail according to the contents displayed on the SERP without clicking. We assume that these distributions have somewhat relationships with the layout of the right-rail results, where results are usually shown in the form of images along with short phrases (entity names). One advantage is that little effort is required for a user to obtain the information and make corresponding judgments, but the disadvantage is that the short phrases can only provide limited information directly due to the text length.

4 Effects of right-rail results

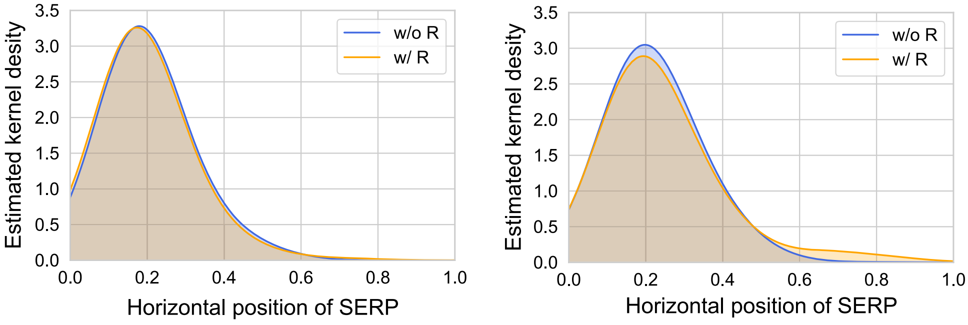
4.1 Examination attention

First, we investigate the attention allocation on the SERPs regarding **RQ1**. As mentioned in Sect. 3, we take the fixation on a result that is no less than 200ms ("valid fixation") as the indicator of examination and dismiss the others, referring to previous work Lorigo et al. (2008), Xie et al. (2017), Li et al. (2018).

Overall patterns Among the 281 SERPs with right-rail results, 44 of them have valid fixations, while only 3 SERPs have clicks on the right rail. Given that the composite right-rail results have lower click necessity, fixation is a more sensitive and suitable examination signal. In general, examining the right-rail results accounts for a modest but substantial proportion (15.7%). We take a deeper look at the examination sequences that have valid fixations on the right rail (denoted as "w/ ER"). The way to the extract examination sequence from eye-tracking data is similar to prior research Arguello et al. (2013), Huang et al. (2011). As results, concerning the start position of examining the right rail in a path, the average and median positions are 69% and 73% in the whole path, respectively. Meanwhile, examination sequences that end with the right-rail examination account for 31.8% in the "w/ ER" SERPs. The results suggest that users tend to examine the right rail in the latter part of their examination. Only one block of results in the right rail would be examined at most of the time (63.64%). Various patterns occur when people examine multiple result blocks in the right rail. One is examining the results of the left and right rail in two separate continuous sequences. The other is examining result items of two sides in an interwoven way. In our study, the former pattern accounts for the majority (62.5%), but due to a high-level of data sparsity and individual differences, we cannot infer a specific examination order in this work while leaving it for the future work.

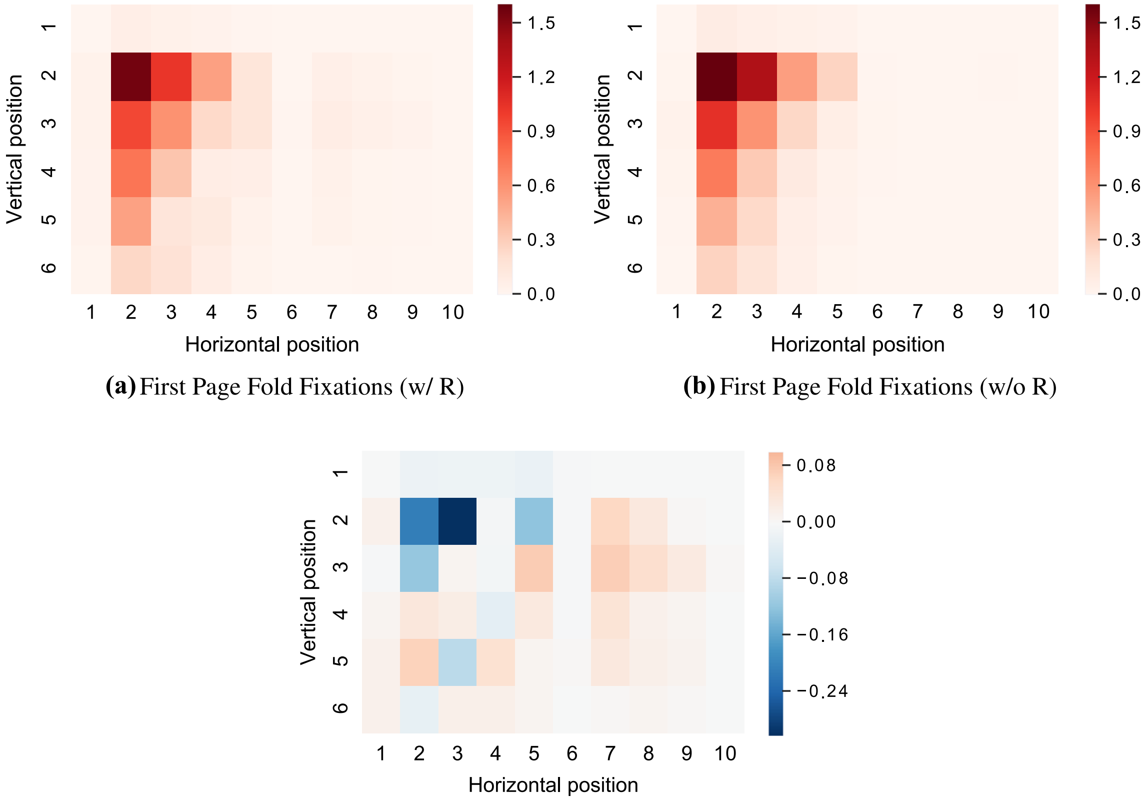
Positional patterns As mentioned in the previous sections, the results of modern SERPs are displayed in more complex layouts. Different results are show with different heights on the SERP, so we use the absolute position instead of the result segmentation in this part. Note that the results in the right rail are always displayed within the first fold of SERP. Thus, we focus on the examination attention distribution within the first-page fold, which commonly contains about three results in the left rail. In Fig. 5, we plot the horizontal distributions of users' first and overall fixations. As for the first fixation position, there is no significant difference between the two distributions (two-sample K–S test, $p = 0.17$), and the peaks both occur in the left half of the page (Fig. 5a). It indicates that users are still used to starting examination from the results in the left part regardless of the existence of right-rail results. In particular, the first examination always starts from the left-top results considering vertical positions of the first arrival. It is also consistent with the former analysis of users' examination sequence that they tend to examine the right rail in the latter part of examination. However, as for the overall fixations, we observe a significant difference between the horizontal distributions (two-sample K–S test, $p = 0.003$). As shown in

Fig. 5 Distribution of “first arrival”/“overall” horizontal position in two experimental conditions. The x-axis denotes the normalized horizontal position of SERP, and the y-axis denotes the estimated kernel density



(a) First Arrival

(b) Overall Fixation



(a) First Page Fold Fixations (w/ R)

(b) First Page Fold Fixations (w/o R)

(c) First Page Fold Fixations (Diff)

Fig. 6 **a**, **b** Are the distributions of fixation on “w/ R” and “w/o R” SERPs, respectively. **c** Is the difference between fixation distributions (“w/ R” - “w/o R”). The x-axis (y-axis) denotes the horizontal (vertical) position in the grid unit

Fig. 5b, with results in the right rail, another small peak occurs in the right half and the corresponding left peak appears to drop a little.

In Fig. 6, we divide the page into six rows and ten columns. The height and width of each grid are both $136.6px$. The attention in each grid is represented by the number of valid examination fixations on the grid and normalized by the number of SERPs under the corresponding experimental condition. In general, we can observe an F-shape pattern of attention in both conditions (Fig. 6a, b). It suggests that users still pay great attention to the left-top results even in the non-linear layout. Since the first row mainly contains the query input box and some query suggestions, it is reasonable that this row draws little attention. However, there do exist some differences between these two distributions. To further clarify the difference, we subtract the attention distributions, i.e., “w/ R” attention distribution - “w/o R” attention distribution, as shown in Fig. 6c. We observe that along with more attention paid to the right side, attention paid to the parallel grids in the left part decreases, especially for the top results. With the results go deeper in the vertical dimension, the difference shrinks.

Table 3 Details of behavioral measures used in the experiment

Measure	Description
query time	Total time (s) of the query
SERP time	Time (s) on the SERP
avg rank (E)	Average rank of fixated results
max rank (E)	Max rank of fixated results
#results left (E)	Number of fixated results in the left rail
#results all (E)	Number of fixated results on the SERP
avg fixation time	Average of fixation time (s) on a result
#revisit (E)	Number of revisiting fixations
#direction change (E)	Number of changes of vertical directions during examination
#results left (C)	Number of clicked results in the left rail
#results all (C)	Number of clicked results on the SERP
TTFC	Time (s) from query to first click
LCTE	Time (s) from last click to ending the query
avg rank (C)	Average rank of clicked results
max rank (C)	Max rank of clicked results
avg interval time (C)	Average time (s) between clicks
#revisit (C)	Number of revisiting clicks
P(C|E)	Probability of clicking on the fixated results
#results left (H)	Number of hovered results in the left rail
#results all (H)	Number of hovered results on the SERP
TTFH	Time (s) from query to first hover
LHTE	Time (s) from last hover to ending the query
avg rank (H)	Average rank of hovered results
max rank (H)	Max rank of hovered results
avg interval time (H)	Average time (s) between hovers
avg hover time	Average time (s) of hovering on a result
#revisit (H)	Number of revisiting hovers
P(H|E)	Probability of hovering on the fixated results
#queries	Number of queries in a session
requery ratio	Ratio of reformulated queries in a session
reform init q	Whether to reformulate the initial query
session time	Total time (s) of a session
SERP time (session)	Time (s) spent on SERPs in a session
#queries w/o C	Number of queries without click
#queries w/o H	Number of queries without hover
#avg requery words	Average number of words in the reformulated queries within a session
#requery voc	Total number of unique words in the reformulated queries within a session

4.2 Search process, satisfaction, and workload

To address **RQ2**, we generate behavioral measures using the eye-tracking and logged data in the user study. Furthermore, we analyze the effects of right-rail results on these behavior measures and users' feedback of satisfaction and workload quantitatively.

Measures Based on the collected eye movements and interactive behaviors during the search process, we compute a variety of query-level and session-level behavioral measures. Table 3 gives the details of each measure. The fixation ($\geq$ 200ms) on a result is considered as an examination signal and the corresponding measures are denoted by "E". Besides, clicks and hovers ($\geq$ 300ms), denoted by "C" and "H", respectively, are also significant measures that represent users' interactions Navalpakkam et al. (2013), Chen et al. (2017). For each type of signal ("E", "C", and "H"), measures related to the numbers and ranks of results as well as the interaction speed are considered. Furthermore, we combine these signals by calculating the conditional probabilities to represent further inspection on a result, e.g., the probability of clicking on the results examined by eye-fixation (P(C|E)). We also include the dwell time measures, e.g., the time spent on the query or the corresponding SERP. In total, we obtain 28 query-level measures. As for the session-level measures, besides the basic dwell time ones (e.g., total time spent on the session, and time on the SERPs in a session), we look at the number of queries in a session and the query reformation strategies. Finally, 9 session-level measures are utilized.

To better understand the underlying aspects of the search process captured by these behavioral measures, we first inspect them via factor analysis. In particular, we conduct a Principal Component Analysis (PCA) using the standardized measures of all sessions and find that the first five components (C0–C4) can explain 67% of the variance. Figure 7 shows the loadings of each behavioral measure onto the five components. Since C3 is mostly overlapped with C4, we merge these two components and finally propose four main factors. **Factor 1** (C0) contains measures related to the ranks and the numbers of examined results and hence is related to users' engagement as well as examination interactions with the SERP. **Factor 2** (C1) is related to session-level efforts and indicates the lack of interactions with the initial SERP. For example, the number of the queries without click/hover and query reformation related measures are in this factor. We assume that when a user has few interactions with the SERP and puts efforts into reformulating queries, the current query might not be successful for her. Note that "the number of clicked results in both/left sides" (denoted by "#results all (C)" and "#results left (C)") also shows an outstanding loading on this factor, but the sign is contrary to that of other measures, so it is also consistent with the above interpretations. **Factor 3** (C2) represents the interaction speed. Time-based measures are included, e.g., "avg intervals (H/C)", "query time", "last click/hover to end (LCTE/LHTE)". **Factor 4** (C3 & C4) includes many revisit behaviors and measures representing further inspection beyond eye-fixation. Thus, it 4 reflects users' comparison among results and cautiousness for further examination.

In addition to the behavior measures, we consider the satisfaction scores of the initial query and the search session as measures for search outcomes. The perceived workload is measured by the NASA TLX Hart and Staveland (1988), containing the levels of mental demand, physical demand, temporal demand, failure, effort, and frustration. In total, we obtain 45-dimensional measures in this part (i.e., 37 behavioral measures and 8 user feedback).

Statistical tests The Mann–Whitney U test Mann et al. (1947), instead of t-test, is applied for statistical test, since most of the measures have non-normal distributions according to the K–S test. Moreover, as a treatment of the multiple comparison problem Fuhr (2018), we conduct the Benjamini–Hochberg procedure Benjamini and Hochberg (1995) in the following experiments, controlling the FDR (False Discovery Rate) at the level of 0.05. In the following analysis, we show the average values of measures, along with the initial *p value* and the significant level controlled by the Benjamini–Hochberg procedure (marked by $*$). Besides the plain Mann–Whitney U Test, we conduct two additional two-way ANOVA(s) to involve the factors of the user and task type, respectively. Note that the main factor of our work is the right rail and we find that the multi-factorial tests reflect consistent results regarding the main factor. Therefore, we make the following analysis based on the results given by the Mann–Whitney U Test.

Regarding **RQ2**, we consider whether to display results in the right rail (i.e., "w/ R" and "w/o R") and whether a user examines the right rail (i.e., "w/ ER" and "w/o ER") as the variables, respectively. In particular, 11 participants have never examined the results in the right rail and 4 tasks that have no valid fixations on the right rail in our user study and thus we filtered out the sessions of these users and tasks when studying the latter variable. Table 4 shows the sample sizes and the results of a post-hoc statistical power analysis using the G*Power 3 program Faul et al. (2007). We calculate the statistical power ($1-\beta$) of the Mann–Whitney U test at a two-tailed significance level of $\alpha = 0.05$. Different effect size parameters measured in Cohen's d Cohen (2013), where 0.2, 0.5, and 0.8 are defined as small, medium, and large effects. The results of post-hoc power analysis indicate that our study can detect medium and large effects in both settings effectively, referring to the conventional threshold of statistical power ($1-\beta \geq 0.8$).

Effects of displaying right-rail results We conduct the Mann–Whitney U test on behavioral, satisfaction, and workload measures. As a result, we do not observe any significant differences in behavioral measures or satisfaction measures. Meanwhile, there is some difference in the workload measure of physical demand, where a higher level of physical demand is reported in the "w/ R" condition ("w/ R" = 1.680, "w/o R" = 0.152, $p = 0.010$), but the difference is not significant considering the FDR-controlling. In summary, the display of right-rail results has a limited influence on user behaviors, search satisfaction, or workloads in our study. The result is reasonable considering that about 1/3 of participants have never examined the right rail in our study, though all of the participants have reported using search engines in the desktop environment frequently.

Effects of examining right-rail results Similarly, we conduct the Mann–Whitney U test on behavioral, satisfaction, and workload measures and then correct the significance via FDR-controlling (marked by $*$). As shown in Table 5, we

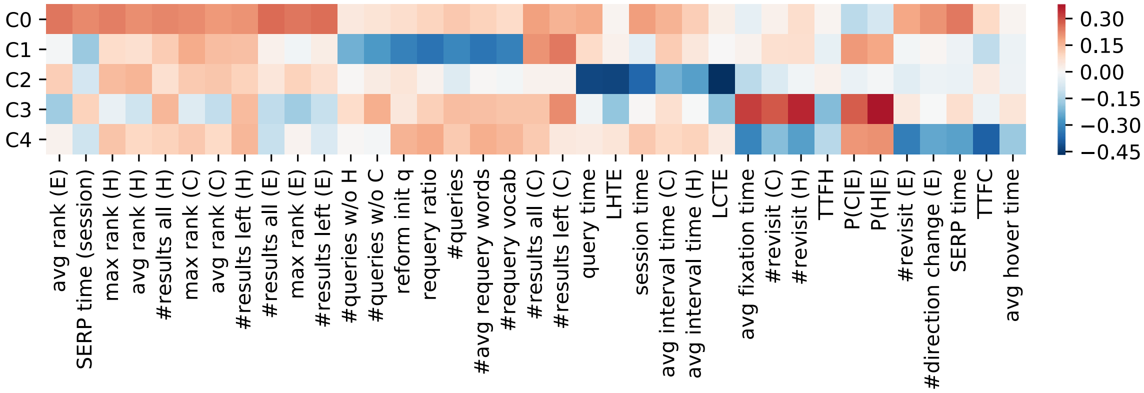


Fig. 7 Clusters of behavioral measures given by PCA. The colors represent the values of loadings

Table 4 Sample size of each condition (i.e., "w/ R", "w/o R", "w/ ER", and "w/o ER") and the results after post-hoc statistical power analysis

Condition	Sample Size		Effect Size d		
	w/	w/o	$d = 0.2$	$d = 0.5$	$d = 0.8$
Right Rail (R)	281	286	0.75	0.99	1.00
Examination (ER)	44	250	0.33	0.90	0.99

Table 5 Differences in search behavior and satisfaction measures

	FID	w/ ER	w/o ER	p_{init}	Sig.
Behavior measure					
#Results_all$_E$	1	7.568	5.296	$2e-5$	$*$
#Results_left$_E$	1	5.750	4.764	0.024	–
SERP_time_session	1	42.70	26.87	$5e-6$	$*$
Query_time_total	3	90.76	75.08	0.039	–
SERP_time_query	4	33.00	19.53	$5e-7$	$*$
#Revisit$_E$	4	4.295	2.360	$2e-4$	$*$
#Revisit$_H$	4	1.614	1.048	0.012	–
#Direction_change$_E$	4	4.886	2.604	$5e-6$	$*$
TTFC	4	14.56	7.044	$8e-4$	$*$
$P(H\|E)$	4	0.3565	0.4284	0.023	–
$P(C\|E)$	4	0.2336	0.3623	$6e-5$	$*$
Satisfaction Measure					
Satisfaction of session		3.432	3.912	$6e-4$	$*$
Satisfaction of query		3.386	3.832	0.013	–

Average valuse are reported. FID denotes the factor ID of the behavioral measure. p_{init} is the initial significant level given by the Mann–Whitney U test and "$*$" denotes that the difference is significant after conducting the Benjamini–Hochberg procedure

can observe significant differences in both behavioral and satisfaction measures.

In terms of search behavior, the measures that are affected significantly can be categorized into two main factors, i.e., **Factor 1**, which is related to users' engagement on the SERP, and **Factor 4**, which reflects comparison among results and carefulness for further examinations. In general, users are more engaged in examining the SERP when they have fixations on the right-rail results. For instance, the time spent on examining SERPs increases in both session-level and query-level measures. Naturally, users examine more results in total if examining the right rail, while it is also interesting that the number of examined left-rail results increases to some degree, though the difference is not significant after the FDR-controlling. Meanwhile, users appear to make more comparisons among results in the left rail according to the increase of revisit fixations and direction changes. Among "fixation", "hover", and "click", we consider that "click" requires the most effort, while "fixation" requires the least. Thus the differences in "P(C|E)" and "TTFC" suggest that users are more careful before leaving the SERP for further examination in the "w/ ER" situation. In summary, when users examine the right-rail results, they also have more interactions with results in the left rail and appear to be more stick to the SERP for information. Furthermore, these differences indicate that users seem to be struggling more during the search in the "w/ ER" situation.

In terms of search satisfaction and workload, we observe some impacts on satisfaction measures, while no significant impacts on self-reported workloads. Users appear to be less satisfied when examining the right rail, which seems a bit surprising. We explain the decrease from the following perspectives. For one thing, the behavioral measures in Table 5 indicate that users put more effort in examining the SERP and appear to be struggling more in the "w/ ER" situation though the differences in explicit workload feedback are not significant. For another, considering the relevance distribution of right-rail results (a larger proportion of irrelevant results), we assume that users might feel depressed when they examine the results that are less useful to complete the task.

Discussion Comparing with previous research Bota et al. (2016), we observe similar effects of displaying results in the rail on search behavior. In general, whether to show results in the right rail has little significant influence on behavioral measures. However, facilitated by the eye-tracking study, we could further observe several significant differences in user behavior when users examine the right rail results, especially in engagement and carefulness measures (i.e., Factor 1 and Factor 4). It is worth mentioning that users reported significantly different workload measures, including mental demand, temporal demand, and effort in the study of Bota et al. (2016). However, few significant effects have been observed on workload measures in our experiments. We think that the different findings might be related to the interface of the experimental SERP. Recall that the results in the left rail in Bota et al. study were simple organic results, while the right rail comprised a composite entity card, involving images and textual information. In that case, the right-rail results would be much more complicated than the left-rail results, and thus more examination effort might be required. On the contrary, the left rail also contains vertical results in our study, which is also consistent with the realistic interface of commercial search engines. When the results of both sides become similarly complicated in presentation forms, the display of right rail results might not increase the perceived workload on a large scale. Furthermore, we found impacts on satisfaction when users examine the right rail results.

4.3 Summary

To sum up, fixation is a more sensitive signal of examination than click in our study. Based on the fixation data, we find that the top results in the left rail are still the main focus of examination on both linear and non-linear SERPs. However, right-rail results will indeed distract some attention and cause a decrease in terms of the attention paid to the top-left items. In particular, users tend to start examining SERPs from the top-left results and examine the right rails in the latter stage of the SERP examination.

Furthermore, we investigate whether displaying or examining the right rail would affect the quantitative measures for the search process, satisfaction, and workload. As a result, the display of right-rail results generally has little influence on the search process, perceived satisfaction, or workload. However, we observe differences in search behavior and satisfaction when users examine the right-rail results. Specifically, users seem to put more effort into interacting with the results on the SERP and be more careful during the examination. Although there is rarely any difference in explicit workload measures, there exists a non-trivial decrease in search satisfaction.

5 Modeling examination behavior on complex SERPs

According to the analysis above, there are differences in both search behavior and search satisfaction when users examine the right-rail results. Therefore, we attempt to predict and interpret the examination behavior on the SERP regarding **RQ3**.

5.1 Data, features, and models

We aim to build a model to predict whether a user u will examine the right rail on a SERP s. Based on the sessions with right-rail results (“w/ R”), we filter out the sessions of the 11 users who have never examined the right rail but maintain the sessions of all the 20 tasks in prediction, because we consider the attributes of the search task itself as influential factors as well. In summary, we obtain 181 SERPs, among which 44 ones have examinations on the right-rail results.

We extract behavioral features (denoted as **BF**) and static features (denoted as **SF**) for each impression of a SERP s. The definitions of these features are shown in Table 6. As for the behavioral features, we select the ones that show significant differences when examining the right-rail results in Table 5. Considering the cost of collecting eye-fixation features, we only use the dwell time and click features, which are much easier to collect for search engines. As for the static features, we utilize the usefulness, attractiveness of left-rail results and the external relevance annotations of right-rail results to model the attributes of result items. Particularly, we do not use usefulness or attractiveness of right-rail results, because if a user gives a usefulness or attractiveness score of the right-rail result, it directly indicates that she has examined that. Besides, we involve some other features of the interface, including the number of right rail results and whether an entity graph is shown in the right rail.

We treat it as a binary classification task. Given the imbalanced label distribution, we evaluate it with AUC (Area Under Curve). The prediction is conducted using fivefold cross-validation, and we report the average performance on the test folds. We have trained several supervised learning methods, including support vector machine (SVM), decision tree, random forest, and gradient boosting decision tree (GBDT). These methods are implemented by sklearn and most of the hyper-parameters are set as default. Specifically, to avoid over-fitting, we use the linear kernel in SVM and set the “max_depth” of each tree up to 2. We did not use a complex neural model in this paper, considering the limited data and its lack of interpretability.

Table 6 Features used in the prediction experiments

Feature group	Notion	Description
Static (Left)	Avg_U_L	Average value of usefulness of left-rail results
	Avg_A_L	Average value of attractiveness of left-rail results
Static (Right)	Avg_R_R	Average value of relevance of right-rail results
	Max_R_R	Maximum value of relevance of right-rail results
	# Results_R	Number of results in the right rail
	Graph	Whether there is an entity graph in the right rail
Behavioral	SERP_T_S	Time (s) spent on the SERPs in the session
	SERP_T_Q	Time (s) spent on the SERP in the initial query
	TTFC	Time (s) to first click

Table 7 Prediction performance (AUC) on different feature groups ("$\triangledown/\triangledown\triangledown$" indicates that the performance is significantly different from that of "SF + BF" at $p < 0.05/0.01$.)

	Random	SF	BF	SF+ BF
SVM	$0.5284^{\triangledown\triangledown}$	$0.5852^{\triangledown\triangledown}$	**0.7596**	0.7453
Decision Tree	$0.5284^{\triangledown\triangledown}$	$0.6192^{\triangledown\triangledown}$	0.6767	**0.7064**
GBDT	$0.5284^{\triangledown\triangledown}$	$0.6418^{\triangledown\triangledown}$	$0.7170^{\triangledown}$	**0.7709**
Random Forest	$0.5284^{\triangledown\triangledown}$	$0.6617^{\triangledown}$	0.7675	**0.7742**

Bold reflects the best performance regarding each classifier

Table 8 Feature importance in prediction models and the pearson's correlation with the examination of the right rail

	Max_R_R	Avg_U_L	Avg_A_L
Feature importance	0.075	0.063	0.045
Pearson's correlation	0.2428	−0.2124	−0.1488
p-value	0.001	0.004	0.045

The "p-value" indicates the significant level of the correlation

5.2 Result analysis

Performance comparison Results are shown in Table 7. In general, different methods reveal similar findings when comparing these feature groups. Both static and behavioral features contribute to predicting whether a user will examine the right-rail results compared with a random classifier. The behavioral features are more effective in prediction than the static ones, even though the "BF" group includes only three-dimensional features. In most learning methods (expect SVM), combining two feature groups achieves the best performance. Meanwhile, the performance of using "BF" is quite close to that of using "BF + SF". The difference in performance between these two feature groups is not significant for most of the classifiers. The results suggest that we can utilize the behavioral features efficiently, which are much cheaper to collect than eye movements and explicit user feedback (e.g., attractiveness and usefulness assessments), to predict the examination of right-rails inferred from eye-tracking.

Feature analysis Furthermore, we attempt to interpret the prediction model by analyzing the roles of different features to explain why users will examine the right rail. Comparing the learning methods in Table 7, the random forest classifier achieved the overall best performance and has good interpretability. Therefore, the following analysis is based on the random forest classifier. Specifically, we analyze the feature importance and the feature correlation with the right-rail examination. We sort the features according to their feature importance in the model trained on "SF + BF" (average value across 5 folds). As expected, the top-3 features are all behavioral features. However, in this part, we pay more attention to the static features for interpretation. Table 8 lists the following top-3 static features along with their feature importance scores. Besides, we calculate Pearson's correlation between the feature and examination. As shown in Table 8, these features have distinct influences. The max relevance of the right rail is a positive factor, while the average usefulness and attractiveness of the left rail are negative ones. The results suggest that users pay more attention to the right rail when it contains some highly relevant items. Meanwhile, the low usefulness level of the left rail will otherwise encourage users to examine the other side of the page. Compared with visual attractiveness, relevance or usefulness are still more influential.

User-independent analysis Besides using behavior, usefulness, and attractiveness as features to predict whether a user examines the right rail on a single SERP, we conduct a user-independent analysis. In this part, we calculate the probabilities of examining the right rail on each SERP by aggregating the labels across users, and we obtain 20 examination probabilities in total. As for features, we only consider the static features and replace the features of usefulness and attractiveness with snippet relevance and click necessity. Instead of prediction, we conduct a correlation analysis for each feature. Table 9 gives the top-3 features ranked according to the absolute value of Pearson's corre-

Table 9 Pearson's correlation coefficients with examination probabilities

	Max_R_R	Graph	Max_R_L
Pearson's correlation	0.6518	0.4974	−0.4352
p-value	0.002	0.026	0.055

The "p-value" indicates the significant level of the correlation

lation coefficients. The max relevance of right-rail results is highly correlated with examination positively. Meanwhile, users are more likely to examine the right side of the SERP especially when the left rail is less relevant. The results are consistent with the analysis in the prediction experiment. Besides, the display of the entity graph also has a positive correlation with the examination. We think that the entity graph might be a visually appealing factor.

5.3 Summary

Regarding **RQ3**, we build prediction models using behavioral and static features and make further interpretations for the prediction models. In summary, the prediction results suggest that the logged behavioral features help effectively predict the examination of the right rail. Meanwhile, the correlation between static features and user examination indicates that the most relevant result in the right rail and the usefulness/relevance of the left part are two main factors. A highly relevant result on the right or a less useful/relevant left rail will encourage the right-rail examination.

6 Conclusions and future work

In this paper, acknowledging that modern SERPs show results in a non-linear complex layout, we focus on investigating the effects of results in the right rail. We conduct an eye-tracking user study to collect rich user behavioral signals as well as explicit user feedback. Based on the collected data, we analyze how the right-rail results influence the allocation of examination attention (**RQ1**), their effects on the search behavior, satisfaction, and perceived workload (**RQ2**). We also build a model to predict when a user will examine the right-rail results and make further explanations by correlation analysis (**RQ3**).

Several interesting findings are made. (1) Despite the non-linear layout and the right-rail results, examination attention is mainly allocated to the left-top results, and users tend to examine the right rail in the latter stage of examining the SERP. However, the right rail still influences the attention flow. It distracts some attention from the left-rail results, especially the top-ranked ones. (2) In general, the existence of blocked right-rail results has little influence in terms of search behavior, reported satisfaction, or workload. However, users appear to have more interactions with results on the SERP and be more careful before conducting further inspection when they examine the right rail. Meanwhile, search satisfaction might decrease in this situation. (3) Behavioral features (e.g., dwell time) can benefit in predicting whether to examine the right rail or not. (4) Users pay more attention to relevance or usefulness during the search, and as a result, a highly relevant right-rail item or deficiency in the left part will attract users' examination of the right rail.

The findings of this work can be further applied to the related research areas, such as evaluation and interface optimization. For instance, although users still tend to focus on the results in the left rail, significant differences occur in both search process and satisfaction when they examine the right rail. Thus, the evaluation mechanism should change accordingly. Given this, understanding and modeling users' examination of right-rail results will also be helpful.

As with any research, there exist some potential limitations of our work which we would like to list as future directions. First, the number of participants is limited, as in most user studies, especially an eye-tracking one. Second, although we include a warm-up task at the beginning of the user study, participants may still misunderstand the experimental procedure and conduct some misoperation, sometimes. We believe that better warm-up training will help. Third, as an attempt to understand users' search process in the current web search environment, we utilize the results returned from the search engine without much manipulation (e.g., manipulating result relevance, vertical type, etc.), and involve different types of search tasks to make the experimental setting close to a natural one. In the future, we plan to explore more experimental controls, e.g., result relevance, some specific search intents, for a more fine-grained investigation. Moreover, we would like to develop an evaluation mechanism for non-linear SERPs based on the results of user modeling.[1, 2]

Funding This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011), Beijing Academy of Artificial Intelligence (BAAI) and Tsinghua University Guoqiang Research Institute.

Availability of data and materials The data set is available at Github[1].

Code availability The code of the chrome extension developed for our user study is available at Github[2].

[1] https://github.com/ThuYShao/UserStudySERPDataset.git.

[2] https://github.com/ThuYShao/ChromeExtension.git.

Declarations

Conflict of interest The opinions expressed in this publication are those of the authors. We have no conflicts of interest to declare that are relevant to the content of this article.

Informed consent All of the participants in the user study have signed the informed consent before starting the experiment.

References

Agichtein E, Brill E, Dumais S (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 (pp. 19–26)

Arguello J, Capra R (2014) The effects of vertical rank and border on aggregated search coherence and search behavior. In: Proceedings of the 23rd acm international conference on conference on information and knowledge management (pp. 539–548)

Arguello J, Capra R, Wu W-C (2013) Factors affecting aggregated search coherence and search behavior. In: Proceedings of the 22nd acm international conference on information & knowledge management (pp. 1989–1998)

Arguello J, Choi B(2019) The effects of working memory, perceptual speed, and inhibition in aggregated search. ACM Transactions on Information Systems (TOIS)3731–34

Azzopardi L, Thomas P, Craswell N (2018) Measuring the utility of search engine result pages: an information foraging based measure. In: The 41st international acm sigir conference on research & development in information retrieval (pp. 605–614)

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological)571289–300

Bota H, Zhou K, Jose J M(2016) Playing your cards right: The effect of entity cards on search behaviour and workload. In: Proceedings of the 2016 acm on conference on human information interaction and retrieval (pp. 131–140)

Bron M, Van Gorp J, Nack F, Baltussen LB, de Rijke M(2013) Aggregated search interface preferences in multi-session search tasks. In: Proceedings of the 36th international acm sigir conference on research and development in information retrieval (pp. 123–132)

Chen D, Chen W, Wang H, Chen Z, Yang Q (2012) Beyond ten blue links: enabling user click modeling in federated web search. In: Proceedings of the fifth acm international conference on web search and data mining (pp. 463–472)

Chen J, Mao J, Liu Y, Zhang F, Zhang M, Ma S (2021) Towards a better understanding of query reformulation behavior in web search. In: Proceedings of the web conference 2021 (pp. 743–755)

Chen Y, Zhou K, Liu Y, Zhang M, Ma S (2017) Meta-evaluation of online and offline web search evaluation metrics. In: Proceedings of the 40th international acm sigir conference on research and development in information retrieval (pp. 15–24)

Chuklin A, de Rijke M (2016) Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In: Proceedings of the 25th acm international on conference on information and knowledge management (pp. 175–184)

Cohen J (2013) Statistical power analysis for the behavioral sciences. Academic press

Dumais S T, Buscher G, Cutrell E (2010) Individual differences in gaze patterns for web search. In: Proceedings of the third symposium on information interaction in context (pp. 185–194)

Faul F , Erdfelder E, Lang A -G, Buchner A, (2007) G∗ power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods 392175–191

Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 765378

Fuhr N (2018) Some common mistakes in ir evaluation, and how they can be avoided. In Acm sigir forum 51:32–41

Hart S G, Staveland L E (1988) Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: Advances in psychology (Vol. 52, pp. 139–183). Elsevier

Hotchkiss G, Alston S, Edwards G, (2005) Eye tracking study

Huang J, White R W, Dumais S (2011) No clicks, no problem: using cursor movements to understand and improve search. In: Proceedings of the sigchi conference on human factors in computing systems (pp. 1225–1234)

Lagun D, Hsieh C -H, Webster D, Navalpakkam V (2014) Towards better measurement of attention and satisfaction in mobile search. In: Proceedings of the 37th international acm sigir conference on research & development in information retrieval (pp. 113–122)

Li X, Liu Y, Mao J, He Z, Zhang M, Ma S (2018) Understanding reading attention distribution during relevance judgement. In: Proceedings of the 27th acm international conference on information and knowledge management (pp. 733–742)

Liu Z, Liu Y, Zhou K, Zhang M, Ma S (2015) Influence of vertical result in web search examination. In: Proceedings of the 38th international acm sigir conference on research and development in information retrieval (pp. 193–202)

Liu M, Mao J, Liu Y, Zhang M, Ma S (2019) Investigating cognitive effects in session-level search user satisfaction. In: Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining (pp. 923–931)

Lorigo L, Haridasan M, Brynjarsdóttir H, Xia L, Joachims T, Gay G,Pan B (2008) Eye tracking and online search: Lessons learned and challenges ahead. J Am Soc Inf Sci Technol 5971041–1052

Luo C, Liu Y, Sakai T, Zhang F, Zhang M, Ma S (2017) Evaluating mobile search with height-biased gain. In: Proceedings of the 40th international acm sigir conference on research and development in information retrieval (pp. 435–444)

Mann H B, Whitney D R (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 50–60

Mao J, Liu Y, Kando N, Zhang M, Ma S (2018) How does domain expertise affect users' search interaction and outcome in exploratory search? ACM Trans Inf Syst (TOIS) 3641–30

Mao J, Liu Y, Zhou K, Nie J -Y, Song J, Zhang M, Luo H (2016) When does relevance mean usefulness and user satisfaction in web search? In: Proceedings of the 39th international acm sigir conference on research and development in information retrieval (pp. 463–472)

Mat-Hassan M, Levene M (2005) Associating search and navigation behavior through log analysis. J Am Soc Inf Sci Technol 569913–934

Navalpakkam V, Jentzsch L, Sayres R, Ravi S, Ahmed A, Smola A (2013) Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In: Proceedings of the 22nd international conference on world wide web, pp 953–964

Rele R S, Duchowski A T (2005) Using eye tracking to evaluate alternative search results interfaces. In: Proceedings of the human factors and ergonomics society annual meeting, Vol. 49, pp 1459–1463

Sakai T, Zeng Z (2020) Retrieval evaluation measures that agree with users' serp preferences: Traditional, preference-based, and diversity measures. ACM Trans Inf Syst (TOIS) 3921–35

Shao Y, Liu Y, Zhang F, Zhang M, Ma S (2019) On annotation methodologies for image search evaluation. ACM Trans Inf Syst (TOIS) 3731–32

Silverstein C, Marais H, Henzinger M, Moricz M (1999) Analysis of a very large web search engine query log. In Acm sigir forum, Vol 33, pp 6–12

Sushmita S, Joho H, Lalmas M, Villa R (2010) Factors affecting click-through behavior in aggregated search interfaces. In: Proceedings of the 19th acm international conference on information and knowledge management, pp 519–528

Teevan J, Alvarado C, Ackerman M S, Karger DR (2004) The perfect search engine is not enough: a study of orienteering behavior in directed search. In: Proceedings of the sigchi conference on human factors in computing systems, pp 415–422

Thomas P, Moffat A, Bailey P, Scholer F, Craswell N (2018) Better effectiveness metrics for serps, cards, and rankings. In: Proceedings of the 23rd australasian document computing symposium, pp 1–8

Thomas P, Scholer F, Moffat A (2013) What users do: The eyes have it. In: Asia information retrieval symposium, pp 416–427

Wang C, Liu Y, Zhang M, Ma S, Zheng M, Qian J, Zhang K (2013) Incorporating vertical results into search click models. In: Proceedings of the 36th international acm sigir conference on research and development in information retrieval, pp 503–512

Wu Z, Sanderson M, Cambazoglu BB, Croft WB, Scholer F (2020) Providing direct answers in search results: A study of user behavior. In: Proceedings of the 29th acm international conference on information & knowledge management, pp 1635–1644

Xie X, Liu Y, Wang X, Wang M, Wu Z, Wu Y, Ma S (2017) Investigating examination behavior of image search users. In: Proceedings of the 40th international acm sigir conference on research and development in information retrieval, pp 275–284

Xie X, Mao J, Liu Y, de Rijke M, Shao Y, Ye Z, Ma S (2019) Grid-based evaluation metrics for web image search. In: The world wide web conference, pp 2103–2114

Zhang F, Mao J, Liu Y, Ma W, Zhang M, Ma S (2020) Cascade or recency: Constructing better evaluation metrics for session search. In: Proceedings of the 43rd international acm sigir conference on research and development in information retrieval

Zhang F, Mao J, Liu Y, Xie X, Ma W, Zhang M, Ma S (2020) Models versus satisfaction: Towards a better understanding of evaluation metrics. In: Proceedings of the 43rd international acm sigir conference on research and development in information retrieval, pp 379–388

Zhang R, Xie X, Mao J, Liu Y, Zhang M, Ma S (2021) Constructing a comparison-based click model for web search. In: Proceedings of the web conference 2021, pp 270–283

Zheng Y, Mao J, Liu Y, Sanderson M, Zhang M, Ma S (2020) Investigating examination behavior in mobile search. In: Proceedings of the 13th international conference on web search and data mining, pp 771–779

Zhou K, Cummins R, Lalmas M, Jose JM (2012) Evaluating aggregated search pages. In: Proceedings of the 35th international acm sigir conference on research and development in information retrieval, pp 115–124