Grid-based Evaluation Metrics for Web Image Search

Xie, X.; Mao, J.; Liu, Y.; de Rijke, M.; Shao, Y.; Ye, Z.; Zhang, M.; Ma, S.

DOI
10.1145/3308558.3313514

Publication date
2019

Document Version
Final published version

Published in
The Web Conference 2019

Grid-based Evaluation Metrics for Web Image Search

Xiaohui Xie
BNRist, DCST, Tsinghua University
Beijing, China
xiexh_thu@163.com

Jiaxin Mao
BNRist, DCST, Tsinghua University
Beijing, China
maojiaxin@gmail.com

Yiqun Liu*
BNRist, DCST, Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
derijke@uva.nl

Yunqiu Shao
BNRist, DCST, Tsinghua University
Beijing, China
shaoyunqiu14@gmail.com

Zixin Ye
Beihang University
Beijing, China
zixinye612@gmail.com

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

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

ABSTRACT

Compared to general web search engines, web image search engines display results in a different way. In web image search, results are typically placed in a grid-based manner rather than a sequential result list. In this scenario, users can view results not only in a vertical direction but also in a horizontal direction. Moreover, pagination is usually not (explicitly) supported on image search search engine result pages (SERPs), and users can view results by scrolling down without having to click a "next page" button. These differences lead to different interaction mechanisms and user behavior patterns, which, in turn, create challenges to evaluation metrics that have originally been developed for general web search. While considerable effort has been invested in developing evaluation metrics for general web search, there has been relatively little effort to construct grid-based evaluation metrics.

To inform the development of grid-based evaluation metrics for web image search, we conduct a comprehensive analysis of user behavior so as to uncover how users allocate their attention in a grid-based web image search result interface. We obtain three findings: (1) "Middle bias": Confirming previous studies, we find that image results in the horizontal middle positions may receive more attention from users than those in the leftmost or rightmost positions. (2) "Slower decay": Unlike web search, users' attention does not decrease monotonically or dramatically with the rank position in image search, especially within a row. (3) "Row skipping": Users may ignore particular rows and directly jump to results at some distance. Motivated by these observations, we propose corresponding user behavior assumptions to capture users' search interaction processes and evaluate their search performance. We show how to derive new metrics from these assumptions and demonstrate that they can be adopted to revise traditional list-based metrics like Discounted Cumulative Gain (DCG) and Rank-Biased Precision (RBP). To show the effectiveness of the proposed grid-based metrics, we compare them against a number of list-based metrics in terms of their correlation with user satisfaction. Our experimental results show that the proposed grid-based evaluation metrics better reflect user satisfaction in web image search.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**.

KEYWORDS

Web image search, User behavior, Evaluation metrics

ACM Reference Format:
Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Yunqiu Shao, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Grid-based Evaluation Metrics for Web Image Search. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3308558.3313514

1 INTRODUCTION

Image search has been shown to be very important within web search. Existing work shows that queries with an image search intent are the most popular on mobile phone devices and the second most popular on desktop and tablet devices [27]. In web image search a different type of search result placement is used compared to general web search, which results in differences in interaction mechanisms and user behavior. Let us consider the image search search engine result page (SERP) in Figure 1 to highlight three important differences: (1) An image search engine typically places results on a grid-based panel rather than in a one-dimensional ranked list. As a result, users can view results not only vertically but also horizontally. (2) Users can view results by scrolling down without having to click on the "next-page" button because the image search engine does not have an explicit pagination feature. (3) Instead of a snippet, i.e., a query-dependent abstract of the landing page, an image snapshot is shown together with metadata

*Corresponding author

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313514

about the image, which is typically only available when a cursor hovers on the result.

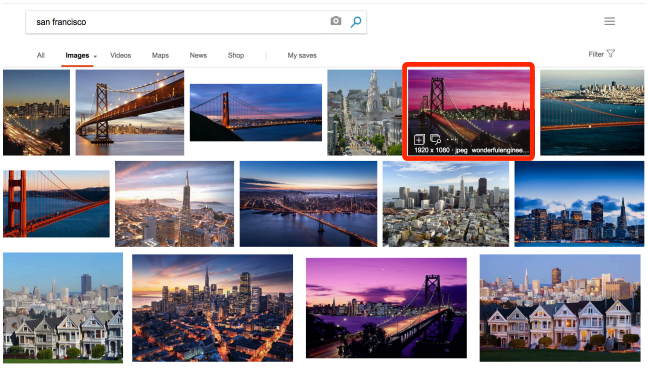

Figure 1: An example SERP from a popular image search engine. For the image that is highlighted using the red box, metadata is displayed when the user hovers over the image.

Evaluation metrics encapsulate assumptions about user behavior [10, 20] and, hence, differences in user behavior should lead to differences in the design of evaluation metrics in image search. Previous work on evaluation metrics [5, 15, 21] focuses on general web search scenarios where results are placed in a list manner. Among the evaluation metrics proposed, Rank-Biased Precision (RBP) [21] assumes that users will examine each result with a persistence probability p from top to bottom; users with a higher value of p are more patient to interact with search results. Discounted Cumulative Gain (DCG) [15] measures the gain of a document based on its position in the result list; the gain is accumulated from the top of the result list to the bottom and is discounted at lower ranks. Although these models work well to assess a result list in general web search, it is not obvious how to adapt them to image search where results are placed in a grid-based manner. Zhang et al. [39] show that the performance of these evaluation metrics is not promising in image search environments in the sense that they do not correlate well with user satisfaction. While the importance of different presentation formats has been recognized [23], there have been very few attempts to construct grid-based evaluation metrics.

As a first step towards designing better evaluation metrics for web image search, we conduct a comprehensive user behavior analysis using data from a lab-based user study so as to obtain a deeper understanding of the underlying user behavior, especially how users allocate their attention.

To summarize, we have three major findings through the analysis:

(1) Similar with the findings in [34], a middle position bias of users attention is observed in the user study data.
(2) The attention of image search users is not discounted monotonically and dramatically along with the rank positions, which means that more attention might not always be allocated to the higher rank positions. Also, the attention allocated to results within a row shows less variance than the attention among different rows.
(3) Users display row-skipping behavior on image SERPs. They may directly jump to results at some distance and ignore particular rows. A two-stage model can be used to depict this process in which users will judge the whole row first, and then decide to skip this row or view the details of results in this row.

Motivated by these observations, we propose corresponding user behavior assumptions to simulate users' interaction processes on SERPs. As shown in [3], evaluation metrics can be generalized as a function of gain and stopping probability, that is the sum over all ranks of the gain (e.g., relevance) accumulated by examining that far, times the probability that this is where the user stops examining the results. The basic idea of our proposed assumptions is to revise the stopping probability by incorporating grid-based position information. We show how we derive new evaluation metrics from these assumptions and how to adopt them to revise well-known list-based metrics.

We conduct extensive experiments to test the proposed assumptions. By using a large-scale commercial image search log, we show that incorporating grid-based features can help user behavior models to better predict the stopping position. We also use data from a field study, in which users' explicit satisfaction feedback and assessors' relevance judgments are available, to measure the performance of the grid-based evaluation metrics. We demonstrate that in image search, existing list-based metrics do not correlate well with user satisfaction while the proposed grid-based evaluation metrics can better reflect user satisfaction.

In summary, we make the following contributions:

- We thoroughly investigate how users allocate their attention on a grid-based interface in image search. We have three major findings of user behavior, i.e., "Middle bias," "Slower decay," and "Row skipping."
- Motivated by our findings on how attention is allocated, we propose corresponding user behavior assumptions to simulate users' search processes. We then derive new grid-based evaluation metrics based on these assumptions.
- We conduct extensive experiments to test the performance of our proposed grid-based evaluation metrics. Experimental results demonstrate that they better reflect user satisfaction and the assumptions behind them are closer to practical user behavior than the assumptions underlying competing models.

2 RELATED WORK

Related work comes in two areas: image search and evaluation metrics.

2.1 Image search

As result placement and interaction mechanisms in image search are different from general web search, user behavior in image search is different from user behavior in general web search. There exists a number of studies on user behavior analysis of image search engines. One line of prior research focuses on characterizing general user behavior based on search logs [2, 12, 28, 32]. Compared with general web search, important differences in user behavior (e.g., shorter queries, a tendency to be more exploratory, and to browse deeper) have been observed. Another line of research investigates more fine-grained user interactions with image SERPs. Xie et al. [34] observe a different browsing model on image SERPs and show a *middle position bias* of users' examination behavior. The observation "Middle bias" in this paper accords with their findings. Also,

interaction behavior such as cursor hovering has been shown to be a valuable additional signal for relevance [22, 35]. User behavior that is unique to image search has motivated various attempts at user behavior modeling that aim to improve the performance of image search engines [14, 35, 37].

Differences in user behavior also have an impact on evaluation. Previous work on evaluation of image search mainly adopts existing list-based evaluation metrics to measure the performance of models developed for image search by simply joining results together [11, 14]. Sanderson [24] introduces evaluation measures used in ImageCLEF, an evaluation forum for cross-language annotation and retrieval of images. However, these metrics still follow those in general web search. Zhang et al. [39] find that existing metrics in web search do not correlate well with user satisfaction in image search. The construction of evaluation metrics that do correlate well with user satisfaction in the context of grid-based interfaces for image search still remains an open question and deserves more attention.

2.2 Evaluation metrics

Evaluation sits at the center of IR research. In order to approximate the system's performance and users' search satisfaction, two components are needed. One is a search result collection labeled with query-dependent relevance levels and the other is a well-designed user model used to simulate the search process [25]. A number of effective evaluation metrics have been designed for general web search [7]. These metrics mainly follow the assumption that users scan ranked results from top to bottom before they stop [9]. One of these, RBP [21], assumes that users examine the $(i+1)$-th result after examining the i-th result with persistence p and will end their examination with probability $1-p$. Järvelin and Kekäläinen [15] propose a metric, DCG, that formalizes user gain from a result list as a discounting process. Besides considering the position impact, Expected Reciprocal Rank (ERR) [5] takes result relevance into consideration and defines the probability that a user is satisfied with a document to be related with relevance of the document. More sophisticated measures have been developed recently. Zhang et al. [38] try to model the search process based on upper limits for both benefit and cost, and propose a Bejeweled Player Model. Also, Wicaksono and Moffat [30] provide a detailed discussion of continuation probabilities (e.g., the persistence p in RBP) in user behavior models that underlie evaluation metrics.

In information retrieval, user satisfaction can be understood as the fulfillment of a specified desire or goal [16]. Satisfaction can be considered as the golden standard in search performance evaluation and is used to reflect users' actual feelings about the system [1, 13]. Correlation with actual user satisfaction is often taken to be the ultimate test for newly proposed evaluation metrics. Indeed, there exists a number of studies investigating different evaluation methods and the correlation between these methods and satisfaction [6, 19, 20, 26]. In this paper, we follow the same principle and also measure the performance of the proposed evaluation metrics by considering their correlation with actual user satisfaction.

What we add on top of the work discussed above is the following. List-based metrics have shown their effectiveness in estimating users' search satisfaction and measuring the performance of general web search engines – but they are list-based. However, in image search a grid-based result placement is adopted. We show that considering grid-based position information as part of the design of evaluation metrics can be beneficial. No previous research has investigated grid-based evaluation metrics for web image search.

3 IMAGE SEARCH USER BEHAVIOR PATTERNS

In order to gain a better understanding of user behavior in image search we examine the attention allocation mechanisms of search users in image search. The findings of this examination will help us to formulate an image search user model that will underlie our proposed grid-based evaluation metric.

We use two publicly available datasets, of image search and web search respectively, in this paper. The image search dataset has been created using data collected in a lab-based user study in image search scenarios [34]. A total of 40 participants have been recruited to complete 20 image search tasks in this study. A Tobbi eye-tracker with default settings has been used to record the examination behavior of participants; the participants' fixation points and fixation dwell time were recorded and certain image being examined was recorded by the built-in algorithms. The general web search dataset has been created using data collected in another user study conducted in general web search scenarios [18]. This dataset involves 32 participants who have been recruited to complete 30 web search tasks. Participants' fixation points on general web SERPs were recorded used the eye-tracker with the same settings and built-in algorithms as in the first dataset described above. Based on these two datasets, we cannot only investigate examination behavior in image search but compare image search with general web search.

We obtain three major findings of user examination behavior on image SERPs. They are "Middle bias," "Slower decay," and "Row skipping." The first one ("Middle bias") is mainly column-based and share the same observations with [34]. Starting from reviewing this finding, we introduce two new observations ("Slower decay" and "Row skipping") which are mainly row-based.

3.1 "Middle bias"

In image search, results are placed in a grid-like manner. Hence, users cannot only examine results vertically, as in web search, but also horizontally, within a row. It is important to investigate how users allocate their attention within a row. For the first dataset, similar to [34], we use the absolute position instead of the border of images to segment SERPs since the number of images in each row may be different (see the SERP example in Figure 1). Each SERP can be equally divided into 5 columns. We then draw a heat map with 10 rows and 5 columns of the distribution of examination durations (averaged over tasks and users); see Figure 2. Here, the examination duration of an image is defined as the dwell time during which a user gazes at the image. Gaze is the externally-observable indicator of human visual attention [17].

By examining the heat map in Figure 2 we re-confirm the observations from [34]: the middle positions in each row receive more attention than other positions, i.e., the leftmost or rightmost positions.

Based on these observations, we propose our first hypothesis:

Hypothesis 1 – Middle bias

Image search results in the middle position may attract more attention from search users than results in the left-most or rightmost position.

Figure 2: Distribution of examination duration (in seconds) in the first 10 rows in image search (0–9); rows are split into 5 columns (0–4).

Hypothesis 1 is not new: Xie et al. [34] already apply a linear mixed model to justify that the middle-position bias is significant statistically. That is, eye gaze behaviors are related to the location of an image within a row, and placing an image in the middle columns has a significant impact on fixation duration. However, they didn't adopt it to construct new image search evaluation metrics.

After Hypothesis 1, which concerns user examination behavior within a row, we introduce two other new observations and hypotheses that concern inter-row examination behavior patterns of image search users.

3.2 "Slower decay"

In image search, users can view results by scrolling down without having to click the "next page" button, which brings less cost to users and results in more exploratory search and deeper browsing depths [32]. We use the eye-tracking user study datasets to investigate how users examine SERPs in image search and general web search.

As shown in [35, 39], different within-row directions have little impact on user behavior modeling in image search. Define the *rank position* in a grid by following the top to bottom and left to right order. We calculate the examination duration for each cell in the grid (in the same way as was used in Section 3.1) and plot the distribution of the top 10 rank positions of image results in Figure 3. For the second dataset, we calculate the examination duration for each result and also plot the duration distribution in Figure 3 for comparison with image search.

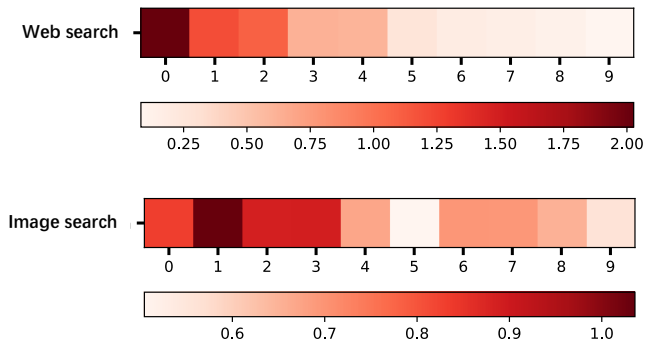

Figure 3: Distribution of examination duration (in seconds) in the first 10 rank positions of general web search and image search.

From Figure 3, the first observation is that users' examination duration does not decrease dramatically with the rank position in image search, especially within the same row (To note, there are five cells within a row.). Also, the difference of values between positions at different rows is smaller than the difference in web search. The second observation is that the change in examination duration in image search is not always monotonic, which is also different from web search. Position 7 (0.694s) receives a longer fixation than position 4 (0.671s) and position 5 (0.505s). In the case of web search, attention decreases in a monotonic way and at a higher speed than in the case of image search.

This leads to our second hypothesis:

Hypothesis 2 – Slower decay

Users' attention does not decrease monotonically and dramatically with the rank position. In the case of image search attention decays at a slower speed than in general web search.

To verify Hypothesis 2, we first take "two distributions in web search and Image search are similar" as the null hypothesis and then we use Pearson's chi-squared test, which is used to determine whether there is a significant difference between the expected distribution and the observed distribution, to determine whether the null hypothesis is true. The result shows that the p-value is less than 0.001. Hence, we can reject the null hypothesis and say that the difference in examination duration distribution between web search and image search in Figure 3 is significant. Also, we define "decay speed" as the result of dividing examination duration in position i by examination duration in position $i + 1$. We calculate the average decay speed based on the data shown in Figure 3. Results show that the average decay speed of image search (1.06) is much lower than of general web search (1.48).

3.3 "Row skipping"

We look deeper into examination sequences of search users using the eye-tracking data. We find that users will not examine every row one-by-one from top to bottom, which means they will skip rows and examine results at some distance. This "Row skipping"

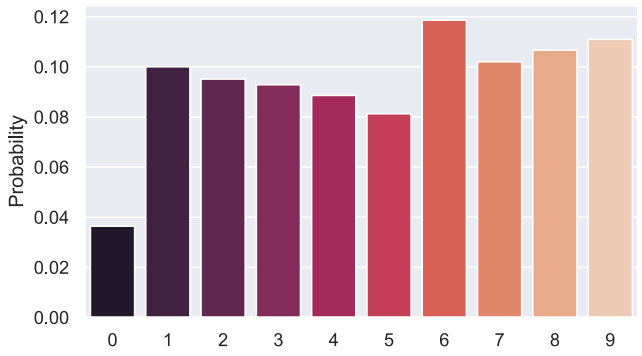

Figure 4: Probability of row-skipping behavior in different rows.

behavior can be formalized as: Right after a user examines results in the i-th row, she/he examines results in the j-th row where $j > i+1$. We define the probability of row-skipping behavior in a certain row (row i) as:

$$P(i) = \frac{S(i)}{S(i) + E(i)}. \tag{1}$$

Here, $E(i)$ is the number of cases where row i is being examined right after row $i-1$ has been examined. And $S(i)$ is the number of cases that users examine results at a row with a row number larger than i after examining row $i-1$. We define "search begin" as the row before row 0. That is, row 0 being skipped means the first examined row is not row 0. We show the probability of row-skipping behavior in the first 10 rows in Figure 4. There exists row-skipping behavior in image search. The highest probability is about 12% in the 6-th row in the first dataset. Also, the row-skipping probability in the 0-th row is much smaller than in later rows, which means users rarely skip the first row in image SERPs. Assuming that participants in a lab-based user study are more patient than users in real-life environments, the probability of row-skipping in real-life can be even higher.

Thus, we propose our third hypothesis:

Hypothesis 3 – Row skipping

Users may ignore particular rows and directly jump to results at some distance.

We take "the frequency of cases that row i being examined after row $i-1$ has been examined (i.e., E(i)) accords with the frequency of all cases that previous examined row is $i-1$ (i.e., S(i)+E(i))" as the null hypothesis. We also perform a chi-squared test and find that the p-value is less than 0.001. Therefore, we can reject the null hypothesis and say that row-skipping behavior does exist in user examination process.

To sum up, we have presented three hypotheses concerning user behavior in image search based on the observations made during eye-tracking user studies. Statistical tests have been conducted to verify the hypotheses and show the significance.

Although our first observation (i.e., "Middle bias") is not new, it has not been adopted in the design of image search evaluation metrics. we devote our attention to it as well as to two other, new observations, since considering both interaction processes in the horizontal direction ("Middle bias") and in the vertical direction ("Slower decay" and "Row skipping") as part of the construction of a grid-based evaluation metrics is beneficial in this two-dimensional environment.

4 GRID-BASED EVALUATION METRICS

In this paper, we construct grid-based evaluation metrics based on the user behavior hypotheses proposed in Section 3. We first introduce a uniform framework from which existing list-based evaluation metrics can be instantiated. We then propose three modeling assumptions motivated by the hypotheses in Section 3. Based on these assumptions, we derive new grid-based metrics by making revisions on the uniform structure.

4.1 Evaluation framework

Given a result set generated in response to a query, we can estimate users' satisfaction based on the relevance score of each query-result pair and a particular user model followed by users when they interact with this result set. Existing list-based evaluation metrics mainly follow an interaction process where users scan ranked results one-by-one from top to bottom before they stop. This interaction process can be regarded as a *cascade model* [9]. Following the cascade assumption, Moffat et al. [20] define a framework that captures a user's expected utility to generalize arbitrary list-based evaluation metrics (M) as:

$$M = \sum_{i=0}^{\infty} W_i R_i, \tag{2}$$

where R_i is the relevance score of the i-th result, and W_i is the metric-specific weight at rank position i. For example, for RBP with persistence probability p, $W_i = (1-p)p^{i-1}$ and for DCG, the metric-specific weight W_i would be $1/\log_2(i+2)$. To note, W_∞ is set to 0 for existing metrics.

Similar to work reported in [3, 38], we construct a uniform framework by considering user continuation and stopping probability. That is, users have a continuation probability C_i at position i to examine the $(i+1)$-th result and with probability S_i they stop at position i and leave this search or issue another query. Thus, S_i can be represented as:

$$S_i = \prod_{j=0}^{i-1} C_j(1 - C_i). \tag{3}$$

As shown in [3], the conditional probability of continuing past the i-th result, i.e., C_i, relates to the metric-specific weight, which can be computed as:

$$C_i = \frac{W_{i+1}}{W_i}. \tag{4}$$

We can transfer the framework mentioned in Eq. 2 to uniform framework depicting user stopping behavior and accumulated gain (relevance) as:

$$\overline{M} = \sum_{i=0}^{\infty}\left(S_i \sum_{j=0}^{i} R_j\right) = \sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1} C_j(1 - C_i) \sum_{j=0}^{i} R_j\right). \tag{5}$$

We refer to $\overline{M}$ as the *total user expected utility*. Next, we show that $\overline{M}$ and M are equivalent (i.e., $\overline{M} \sim M$):

$$\begin{aligned}\overline{M} &= \sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1} C_j(1-C_i)\sum_{j=0}^{i} R_j\right)\\ &= \frac{1}{W_1}\sum_{i=0}^{\infty}\left((W_i - W_{i+1})\sum_{j=0}^{i} R_j\right)\\ &= \frac{1}{W_1}\sum_{i=0}^{\infty}\left(R_i\sum_{j=i}^{\infty}(W_j - W_{j+1})\right)\\ &= \frac{1}{W_1}\sum_{i=0}^{\infty} R_i W_i\\ &\sim M.\end{aligned} \tag{6}$$

The last equivalence holds because W_1 is a constant given a certain evaluation metric. The framework detailed in Eq. 5 can take the user interaction process into consideration more naturally than the framework depicted in Eq. 2 which mainly models the metric-specific weight and obtained gain for each rank. We therefore make revisions on this framework by incorporating grid-based assumptions. For convenience, we use a triple $(i, r(i), c(i))$ to represent the index of an image result. As we discuss in Section 3.2, we predefine the examination order of search users in image search to be from top to bottom and from left to right. Based on this order, we can obtain the rank position i of a certain image which is in the $r(i)$-th row and $c(i)$-th column.

We are now in a position to introduce the grid-based modeling assumptions which are among the contributions of our work. The order in which we propose our assumptions is the same as the order used to present observations of user behavior patterns in Section 3, i.e., "Middle bias" followed by "Slower decay" and "Row skipping."

4.2 Middle bias assumption

The first assumption, named "Middle bias," focuses on the interaction within a single row, i.e., it is column-based. As mentioned in Section 3.1, users have a higher probability to examine results in the middle position. In this paper, we simulate this bias by considering users' continuation examination, in which we increase the stopping probability in the middle position and lower it in the leftmost or the rightmost positions. We assume that users will have a higher probability to finally stop at the middle position within a row. Hence, we can use a column-based function $f(c)$ to modify the stopping probability S_i. For the image at rank position i with the column number $c(i)$, we design the function $f(c(i))$ as follows:

$$f(c(i)) = e^{g(c(i))}, \tag{7}$$

where $g(c(i))$ is a normal distribution with mean μ and standard deviation σ as:

$$g(c(i)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(c(i)-MP-\mu)^2}{2\sigma^2}}, \tag{8}$$

where MP denotes the column index of the middle position in row $r(i)$. We leave explorations of other functions (such as, e.g., a quadratic function) as future work. In a normal distribution, the mean is the central tendency of the distribution; it defines the location of the peak for normal distributions. And the standard deviation is a measure of variability; it defines the width of the normal distribution. Since we simulate users' middle bias in this assumption, we set μ to be a constant number 0 to further simplify the parameter estimation process, which means the "location" of the normal distribution will be right in the middle of the column. Thus, σ is then the only parameter needed to estimate in Eq. 8. Hence, based on the middle bias assumption, the total user expected utility ($\overline{M}$) can be represented as:

$$\overline{M_{MB}} = \sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1} f(c(i))C_j(1-C_i)\sum_{j=0}^{i} R_j\right). \tag{9}$$

4.3 Slower decay assumption

As mentioned in Section 3.2, the "Slower decay" observation shows that users are more patient in image search than in web search. Their attention decreases more slowly, especially on results within a row. Thus, simply adopting existing evaluation metrics, developed for web search, to image search scenarios is not promising. In this paper, we utilize the row information of image results. We assume that users' stopping probability will increase along with the row. Hence, we can revise the stopping probability S_i in the proposed evaluation framework by multiplying S_i with a row-based function $I(r)$. Considering a result at rank position i with row number $r(i)$ and column number $c(i)$, the revised probability of stopping at this result can be computed as:

$$\overline{S_{i,r(i),c(i)}} = I(r(i)) \cdot S_{i,r(i),c(i)}, \tag{10}$$

where $S_{i,r(i),c(i)}$ is the original stopping probability of a certain list-based metric; $I(r)$ is a monotonically increasing function. In this paper, we define $I(r)$ as an exponential function with a base β larger than 1. Then, we can rewrite Eq. 5 as:

$$\overline{M_{SD}} = \sum_{i=0}^{\infty}\left(\prod_{j=0}^{i-1} \beta^{r(i)} C_j(1-C_i)\sum_{j=0}^{i} R_j\right). \tag{11}$$

Adding the parameter β can slow down the speed of decreasing the stopping probability along with rows, since users might still have a relatively high probability of examining results at a lower rank (see Figure 3). Also, the stopping probability of results within a row will multiply the same value according to Eq. 10, which attempts to control the variance between the stopping probability of results in the same row. When $\beta = 1$, Eq. 10 models the stopping probability of existing list-based metrics. We show how different values of β affect the estimated stopping probability distribution in Section 6.

4.4 Row skipping assumption

The third assumption is motivated by the "Row skipping" observation which suggests that users may skip particular rows and jump to results at some distance. In this paper, we model this process by considering a two-stage browsing process. In the first stage, users briefly browse the whole row; we can join image results within a row together to an imaginary "united image". By viewing this "united image," users will make a decision for the second stage where they either skip this row or examine results in this row in detail. We arrive at this two-stage browsing process motivated by a

neuroimaging study [36], which gives important hints about the multistage mechanisms of visual perceptual learning in the brain.

We are now in a position to describe our row skipping evaluation metric ($\overline{M_{RS}}$). We use a parameter γ to depict the probability with which users skip the next row after examining the current row; γ is also a trainable parameter with a value between 0 and 1. Then, the stopping probability of users at rank position i can be computed as follows:

$$S_{i(RS)} = \underbrace{\prod_{k=0}^{r(i)-1}\left((1-\gamma)\prod_{j=S(k)}^{S(k)+N(k)-1}C_j+\gamma\right)}_{\text{two-stage browsing assumption}}\cdot(1-\gamma)\prod_{j=S(r(i))}^{i-1}C_j(1-C_i), \tag{12}$$

where $N(k)$ is the number of images in the k-th row and $S(k)$ is the total number of images before the k-th row. The first part of Eq. 12, before the multiplication sign, depicts the two-stage browsing assumption. We simply assume that with a probability $(1-\gamma)$, users will examine all the image results within this row. Since users stop at row $r(i)$, they will not skip row $r(i)$. Thus, there is the probability $(1-\gamma)$ in the second part after the multiplication sign in Eq. 12.

The row skipping assumption also has an impact on the accumulated gain (i.e., $\sum_{j=0}^{i} R_j$). Since users have a probability γ to skip a certain row, the gain received from this row should be discounted by multiplying $(1-\gamma)$. Hence, the total user expected utility ($\overline{M}$) based on the row skipping assumption can be computed as:

$$\overline{M_{RS}} = \sum_{i=0}^{\infty} S_{i(RS)} \sum_{j=0}^{i}(1-\gamma)R_j. \tag{13}$$

In this section, we have proposed three grid-based assumptions. According to these assumptions, we revise the formula expressing the continuation probability, stopping probability and also the accumulated gain in the uniform evaluation framework (see Eq. 5). To sum up, we modify the stopping probability at different columns, increase the value in the middle position, by considering a normal distribution according to the "Middle bias" assumption. We modify the stopping probability at each rank by increasing the value of the probability along with the rank according to the "Slower decay" assumption. And based on the "Row skipping" assumption, we consider a two-stage browsing process in which users have a skipping probability to ignore a certain row. Thus, the accumulated gain of a certain row is also modified by multiplying the probability that users browse this row.

5 EXPERIMENTAL SETUP

We evaluate the proposed grid-based evaluation metrics using search logs from a commercial image search engine and data from a field study, in which query-level satisfaction feedback and assessors' relevance judgments for query-image pairs are available. Since there is a user behavior model, which depicts the stopping behavior of search users, behind each proposed assumption, we first perform a sanity check, that is, an experiment to test if incorporating grid-based features can help the underlying user behavior model to better predict the stopping position (in terms of mean log-likelihood). As mentioned in Section 2, user satisfaction can be considered as the golden standard in search performance evaluation. In the same way as in [19, 38], we compare our proposed grid-based evaluation metrics against existing list-based metrics in terms of their correlation with user satisfaction to show the effectiveness of proposed grid-based assumptions.

In this section we first introduce the datasets and then describe the design of the two main experiments in this paper.

Table 1: Statistics of the datasets used in our experiments ("#" refers to "number of").

Dataset	#Tasks	#Participants	#Queries	#Sessions
Search log	–	–	82,629	100,000
Field study	555	50	1,212	1,212

5.1 Datasets

Two image search datasets are used to conduct the experiments. Descriptive statistics of these two datasets can be found in Table 1.

The first dataset ("Search log") is randomly sampled from a search log in October 2017 from the Sogou image search engine, which is popular in China. In this dataset, the grid-based information (i.e., row and column number of image results) and user interaction behavior (i.e., click and cursor hovering) are available. We keep query sessions that have at least one click to make sure we can estimate the user's stopping position, since the last clicked rank can be used to approximate the users' actual stopping rank as shown in [3]. The number of search sessions used in this paper is 100K in total, with 80,000+ distinct queries. We split all query sessions into training and test sets at a ratio of 8:2.

The second dataset ("Field study") consists of data collected from a one-month field study, which is publicly available (see [31]). In this field study, participants are asked to provide explicit satisfaction feedback for their search experience. To note, they can decide which query sessions they want to give the explicit feedback on without having to annotate all search sessions and they are also asked to provide a description about the task they conduct when issuing a specific query. Query-level satisfaction scores on a 5 point scale are gathered. Besides user behavior data recorded using a browser extension and explicit feedback from participants, relevance scores of query-image pairs are annotated by assessors on a crowdsourcing platform. Each query-image pair has at least five relevance annotation scores in the range of 0 to 100. We use the average of these annotation scores in our experiment as the label of a certain query-image pair. Also, assessors are recruited to assign a user intent tag to each task (i.e., the "Locate, Learn, Entertain" taxonomy proposed by Xie et al. [33]). Since image search users usually have deeper browsing depths, we test the performance of evaluation metrics at depths of 5, 10 as in [39] and 15 as well. Thus, we keep query sessions in which the number of the last browsing row is not less than 15, which leads to 1,212 query sessions in total in our dataset.

5.2 Experiment 1: Behavior prediction

Experiment 1 is aimed at testing whether the proposed grid-based user behavior assumptions (considering the continuation and stopping behavior in a grid-based interface) are closer to real-life user

Table 2: Grid-search values of hyper-parameters that need to be estimated and the grid-based assumptions these parameters belong to for Experiment 1.

Parameter	Assumption	Grid-search values
p	–	$\{0.1, 0.2, \ldots, 0.9\}$
σ	Middle bias	$\{1, 2, \ldots, 10\}$
β	Slower decay	$\{1.1, 1.2, \ldots, 2.0\}$
γ	Row skipping	$\{0.1, 0.2, \ldots, 0.9\}$

behavior than list-based assumptions. As mentioned in Section 4, the proposed grid-based assumptions revise users' stopping probability at different rank positions by incorporating row and column information. To validate the user behavior assumptions underlying the proposed evaluation metrics, we test the performance of these assumptions on predicting users' actual stopping positions.

We use RBP as our baseline model that naturally takes users' continuation and stopping into consideration. In RBP, a persistence probability p is used to depict users' continuation probability at each rank. Based on formulas introduced in Section 4, we can calculate the stopping probability at different ranks, estimated by RBP as well as by grid-based RBPs with different proposed assumptions (i.e., "MD": Middle bias; "SD": Slower decay; "RS": Row skipping). For example, the stopping probability at rank position $(i, r(i), c(i))$ estimated by grid-based RBP with the "Slower decay" assumption according to Eq. 11 can be computed as:

$$S_i(RBP - SD) = \prod_{j=0}^{i-1} \beta^{r(i)} p(1-p). \tag{14}$$

We regard the last click position to be users' stopping position on SERPs in the same way as in [3]. And we use log-likelihood to show how well the stopping probability distributions estimated by different models approximate the actual user stopping behavior.

We use a grid-search algorithm to estimate the best parameter(s) for each model to minimize the mean log-likelihood of the training data (80%) in our first dataset. We then test the performance of these models with the pre-trained hyper-parameter(s) in the test data (20%). We show the details of the bounds and discretization of the different parameters needed to be estimated using grid-search in Table 2.

5.3 Experiment 2: Correlation with user satisfaction

In Experiment 2, we measure the performance of our grid-based assumptions by testing the correlation between grid-based evaluation metrics, derived from our assumptions, and user satisfaction. We first conduct experiments on RBP-based metrics. We show Pearson's correlation results of RBP with different assumptions (the original list-based and the proposed grid-based assumptions). We also construct a t-statistic to test the significance of the difference between two dependent correlation coefficients [8]. The p-value level is reported if a significant difference is observed. We then look deeper into the effect of different settings of our proposed assumptions (e.g., different starting rows to consider row-skipping assumption, different number of rows of results being modeled in the evaluation metrics). After that, we report results of the grid-based evaluation metrics, under the best settings, based on other list-based prototype metrics (i.e., ERR and DCG). Comparisons are also made between different prototype metrics.

6 RESULTS

We first report the results of Experiment 1, behavior prediction of user behavior models that are based on different grid-based assumptions. Then, in Experiment 2 we show the performance of grid-based evaluation metrics in terms of their correlation with user satisfaction. We compare the parameter selection in different tasks and discuss the optimal settings to perform our proposed grid-based evaluation metrics. Additional comparisons are made between different grid-based evaluation metrics based on different prototype list-based evaluation metrics.

6.1 Evaluation of behavior prediction

Table 3 shows the minimized mean log-likelihood of each user behavior model as well as the value of the best parameters and improvements over the baseline model (RBP). Here, the improvement of the log-likelihood of model A over model B is computed as $\frac{LL(B)-LL(A)}{LL(B)}$. We also perform pairwise t-tests to determine the significance of the observed difference between grid-based models and the baseline model.

Table 3: Outcomes of Experiment 1. Minimized mean Log-likelihood of user behavior models. **: significantly better than the RBP model with p-value < 0.01.

Model	Parameter(s)	Log-likelihood	Improvement
RBP	$p(0.8)$	-0.542	–
RBP-MB	$p(0.8)$, $\sigma(2)$	-0.513^{**}	5.4%
RBP-SD	$p(0.8)$, $\beta(1.9)$	-0.473^{**}	12.7%
RBP-RS	$p(0.5)$, $\gamma(0.2)$	-0.469^{**}	13.5%

Compared against the baseline model (RBP), our grid-based models with the proposed assumptions achieve better performance on behavior prediction, i.e., users' stopping behavior, in terms of mean log-likelihood. Also, all observed differences are significant. The best grid-based model RBP-RS obtains a 13.5% (significant) improvement over the list-based model RBP. Thus, incorporating grid-based information into the construction of a user behavior model is beneficial and results show that search user behavior in a grid-based environment differs from that in a list-based environment.

Compared to the "Slower decay" and "Row skipping" assumptions, both of which help RBP to better predict user stopping behavior, RBP with the "Middle bias" assumption has a smaller improvement over the baseline model on behavior prediction. The reason can be two-fold:

(1) The method used to depict middle position bias of search users may not be optimal. The practical distribution of users' stopping probability within a row may follow more complex distributions. We leave an investigation on methods to more accurately model "Middle bias" behavior as future work.
(2) Users' stopping behavior correlates more with row information than with column information. Thus, the row-based assumptions ("Slower decay" and "Row skipping") achieve better results

than the column-based assumptions ("Middle bias") on behavior prediction.

We also show the value of the best parameters in Table 3. We can see the performance of the baseline model with a fixed continuation probability is not promising, which indicates that in image search users' continuation probability may be affected by other factors like the position information of the current examined result.

When considering the "Middle bias" assumption, the value of the best parameter σ is 2. For a normal distribution, a small standard deviation (σ) produces a distribution that is more tight. Thus, a difference in stopping probability between middle position and other positions is observable.

By incorporating the additional parameter β (when considering the "Slower decay" assumption), we are able to consider the possibility that users' stopping probability will increase along with the row. In this setting, the probability of the results at lower ranks being examined will be higher than in the list-based setting, which might indicate that in image search users have deeper browsing depth (confirming [32]); the stopping probability will decrease slowly.

In RBP with the "Row skipping" assumption, we observe that the probability to ignore certain rows is 0.2. This observation accords with the results shown in Figure 4 while the row-skipping probability in the search log is slightly higher than in the user study data. This may be caused by the fact that participants in a lab-based user study may be more patient due to the phenomenon that their behavior will be recorded. Thus, the probability of row-skipping of real-life users can be higher.

In summary, Experiment 1 has shown that the grid-based assumptions proposed in this paper are closer to natural user behavior than list-based assumptions. User behavior models underlying the grid-based assumptions achieve better performance in predicting real-life user behavior, i.e., users' stopping behavior. The value of the estimated parameters of grid-based assumptions further confirms the observations introduced in Section 3.

6.2 Evaluation of user satisfaction correlation

As explained in Section 5.3, we first consider RBP-based evaluation metrics at top 10 rows, in the same way as in [39]. Table 4 shows the coefficients of Pearson's correlation between RBP-based metrics and user satisfaction. As shown in Figure 4, the row-skipping probability in the 0-th row is much smaller; we also compare the different settings of where we start to perform "Row skipping" in this table.

We can observe from Table 4 that with the help of the proposed grid-based assumptions, RBP-based evaluation metrics can achieve better correlation with user satisfaction than the original RBP that follows the list-based assumption.

Since the optimization target is different from the target in Section 6.1, we fit the best parameters of different evaluation metrics to gain the best correlation with users satisfaction in this experiment. We can observe from Table 3 and Table 4 that the best parameters in these two tasks are slightly different. The reason can be two-fold:

(1) We consider a fixed number of rows in this experiment to calculate the correlation, since a predefined scale of results being measured is required for offline evaluation metrics [15, 21]. However, in Experiment 1, we compute the log-likelihood based on the rank users stop at. Real users may have different depths of browsing due to their search dwell time.
(2) In the field study, participants can freely decide the feedback of which query session to be recorded by the browser extension. Thus, the search intent distribution may be slightly different between the field study data and the search log. This difference has previously been observed by [33]. Different search intents have an impact on user behavior and satisfaction [31].

Table 4: Outcomes of Experiment 2. Pearson's Correlation between RBP-based evaluation metrics (@top 10 rows) and user satisfaction in the field study dataset. "S@n": "Start performing Row skipping assumption at row n". ↑ (↓): the grid-based evaluation metric achieves better (worse) performance than the baseline model RBP. All correlations are significant at the $p < 0.001$ level.

Metric	Parameter(s)	PC coefficient	Performance
RBP	$p(0.7)$	0.333	–
RBP-MB	$p(0.7)$ $\sigma(1)$	0.341	↑
RBP-SD	$p(0.7)$ $\beta(1.2)$	**0.342**	↑
RBP-RS(S@0)	$p(0.8)$ $\gamma(0.1)$	0.322	↓
RBP-RS(S@1)	$p(0.7)$ $\gamma(0.2)$	0.336	↑
RBP-RS(S@2)	$p(0.7)$ $\gamma(0.2)$	0.334	↑
RBP-RS(S@3)	$p(0.7)$ $\gamma(0.2)$	0.333	–

Table 5: Outcomes of Experiment 2. Pearson's Correlation between RBP-based evaluation metrics at different number of rows (top 5, 10 and 15 respectively) and user satisfaction in field study dataset. "S@n" refers to "Start performing row skipping at row n". All correlations are significant at the $p < 0.001$ level.

Metric	Top 5 rows	Top 10 rows	Top 15 rows
RBP	0.331	0.333	0.333
RBP-MD	0.340	0.341	0.341
RBP-SD	0.342	0.342	0.342
RBP-RS(S@1)	0.331	0.336	0.337

We also investigate how the choice of different first rows to which to apply the "Row skipping" assumption affects the performance of the grid-based evaluation metrics. Results are shown in Table 4; they indicate that applying the "Row skipping" assumption at the very beginning is not promising. When we apply the "Row skipping" assumption at the second row (i.e., RBP-RS(S@1)), we observe a better result, with a higher correlation with user satisfaction than list-based RBP. We also show the results of other RBP-RSs with different starting rows (with a row number larger than 1) to apply the "Row-skipping" assumption in Table 4. We find that although these metrics are better than the list-based metric, the improvement of them over the baseline decrease along with the row number of the starting rows. RBP-RS(S@3) has no observable improvement. Thus, it is optimal to consider "Row skipping" starting from the second row. This finding accords with the results plotted in Figure 4, which shows that users rarely skip the first row on an image SERPs.

Since we need to define the number of rows of results being considered in the evaluation metrics before using a certain metric, we also discuss the optimal setting of the row scale. For each query

session, we test the performance of different RBP-based models at the top 5, 10 and 15 rows respectively. Results are shown in Table 5. We have two findings from this table:

(1) When only a small number of rows is considered, grid-based evaluation metrics with the "Row skipping" assumption, which mainly takes row-based information into consideration, cannot achieve improvements over the list-based competitor. However, the column-based evaluation metric (i.e, RBP-MD) is still better than the baseline model: RBP-MD mainly considers user behavior within a row: changes in the number of rows have less effect on it. For the "Row skipping" assumption, the reason that we only observe small differences may be that the stopping probability at a lower rank is too small to be affected. Thus, the improvement over RBP obtained by considering "Row skipping" is achieved mainly from the top rows.
(2) The more rows are being considered in the evaluation metrics, the better the correlation with user satisfaction that can be achieved, for all evaluation metrics (see, e.g., how "Top 10 rows" compares with "Top 5 rows"). However, the difference between "Top 10 rows" and "Top 15 rows" is small which indicates there exists an upper bound on the performance. Hence, considering the annotation expense, we regard "Top 10 rows" as the best setting of the row scale.

Armed with the best settings ("S@1", "Top 10 rows") observed from the experiments conducted on the RBP-based evaluation metrics, we further test the effectiveness of our proposed grid-based assumptions on other list-based prototype metrics. We perform experiments on two other list-based prototype metrics, i.e., DCG and ERR. Recall that DCG is also a position-based model, like RBP. The difference is that the continuation probability of the result at rank i in DCG is rank-dependent; it can be computed as:

$$C_i(DCG) = \frac{W_{i+1}}{W_i} = \frac{\log_2(i+2)}{\log_2(i+3)}. \tag{15}$$

In addition, we consider ERR. Unlike DCG and RBP, the stopping criterion of ERR is affected by the gain (G) of the currently examined result. Following [5], the probability that a user stops at rank i can be represented as:

$$S_i(ERR) = \prod_{j=0}^{i-1}(1-G_j)G_i, \tag{16}$$

where G_i is the gain that correlates with the relevance score of the current result at rank i, which has the following form:

$$G_i(ERR) = \frac{2^r - 1}{2^{r_{\max}}}. \tag{17}$$

where r is the relevance score of the i-th result. ERR and DCG have been used in previous evaluation tasks on image search [35, 39].

We are now in a position to test the performance of our grid-based assumptions on these two evaluation metrics. The results are presented in Table 6. The proposed grid-based assumptions can help ERR and DCG to achieve better correlation with user satisfaction, while an exception is observed (ERR with the MB assumption). All ERR-based evaluation metrics obtain a poor correlation with user satisfaction, confirming a similar result by Zhang et al. [39]. The reason may be that ERR focuses more on the user gain rather than the examined position. As shown in [11, 35], users' judgments about image results depend largely on image attractiveness. Only considering the effect of relevance on user stopping may not be promising. Furthermore, since position information is not explicitly modeled in the stopping probability in ERR, a grid-based version of ERR cannot achieve promising results.

Table 6: Outcomes of Experiment 2. Pearson's Correlation between evaluation metrics (DCG and ERR @top 10 rows) and user satisfaction in field study dataset. "(u@0.9)" refers to "Upper bound of continuation probability is 0.9". All correlations are significant at the $p < 0.001$ level. ‡ (†): the difference is significant comparing to the corresponding list-based metrics at the $p < 0.01$ (0.05) level.

Metric	List-based	MB	SD	RS(S@1)
ERR	0.169	0.152	0.169	0.180
DCG	0.225	0.260‡	0.224	0.295‡
DCG (u@0.9)	0.291	0.308‡	0.305†	0.303

For the DCG-based evaluation metrics, we can observe the expected results that most grid-based DCG metrics perform better than the list-based DCG, demonstrating the effectiveness of our grid-based assumptions. We also observe a similar performance of DCG and DCG-SD. This may be explained by the fact that the original continuation probability, which is shown in Eq. 15, approaches 1 quickly along with the rank which results in a small stopping probability approaching 0. Thus, the parameter β of "Slower decay" assumption has limited effect on the stopping probability. We also consider an upper bound of the continuation probability of DCG-based evaluation metrics. The results are also shown in Table 6 (last row). All grid-based DCG metrics obtain better correlation with user satisfaction than the list-based DCG. Also, setting an upper bound on the continuation probability improves the performance of all DCG-based metrics, which confirms the observation that users' attention decays at a slower speed. Simply adopting assumptions of list-based DCG is not promising in image search scenarios. Importantly, the best parameters of grid-based DCG are almost the same as for RBP shown in Table 4, i.e., $\sigma(1)$, $\beta(1.1)$ and $\gamma(0.2)$, where the different setting for β may be caused by the different continuation probability settings between RBP and DCG. The results shown in Table 6 indicate that the proposed grid-based assumptions help increase the correlation of position-based models of user satisfaction (e.g., RBP and DCG).

In summary, Experiment 2 has shown that the proposed grid-based assumptions can help existing list-based evaluation metrics, especially position-based evaluation metrics (e.g., RBP and DCG), to better reflect user satisfaction. We find that: (1) performing the "Row-skipping" assumption beginning at the second row rather than the first row can help RBP-RS to achieve better performance; and (2) a result grid limited to the top 10 rows in RBP-based evaluation metrics is optimal considering the trade-off between metric performance and annotation cost.

7 CONCLUSION AND FUTURE WORK

In this paper, we have conducted a comprehensive user behavior analysis using data from a lab-based user study so as to understand the attention allocation mechanisms of search users in image search.

We obtain three major findings through our analysis: (1) User attention follows a middle position bias within a row ("Middle bias"). (2) User attention in the case of image search decays more slowly than in general web search ("Slower decay"). (3) Users may skip particular rows and jump to results at some distance ("Row skipping").

We have proposed three grid-based assumptions. Our experimental results show that user behavior models underlying these grid-based assumptions are closer to real-life user behavior. Existing evaluation metrics (e.g., RBP and DCG) can achieve better performance in terms of correlation with user satisfaction by taking grid-based assumptions into consideration.

Our work is the first attempt to construct grid-based evaluation metrics for Web image search. The research outputs of this paper can guide the optimization of image search engines (e.g., in result ranking and UI design) and are also meaningful to inform user behavior modeling in grid-based environments (not only image search but also video search and e-commerce).

Limitations of the proposed grid-based assumptions which may guide future work: (1) The proposed grid-based assumptions mainly consider the effect of the position. It may be beneficial to also take appearance bias (the effect of image attractiveness) into consideration. (2) The way to model grid-based user behavior may not be optimal, e.g., using the normal distribution to simulate the "Middle bias." Methods to encode grid-based user behavior and combine different user behavior assumptions need further investigation [4]. (3) We test the performance of grid-based assumptions on a small group of evaluation metrics only. Experiments conducted on further evaluation metrics are called for. (4) As the effectiveness of evaluation metrics may vary with tasks [29], we will try to investigate the performance of proposed grid-based evaluation metrics across search tasks and intents.

Code

To facilitate reproducibility of our results, we share the code used to run our experiments at https://github.com/THUxiexiaohui/grid-based-evaluation-metrics.

Acknowledgements

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011), The National Key Research and Development Program of China (2018YFC0831700), Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), and the Innovation Center for Artificial Intelligence (ICAI). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 773–774.

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

[3] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 605–614.

[4] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A neural click model for web search. In *WWW 2016: 25th International World Wide Web Conference*. ACM, 531–541.

[5] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *18th ACM conference on Information and knowledge management*. ACM, 621–630.

[6] Aleksandr Chuklin and Maarten de Rijke. 2016. Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In *25th ACM Conference on Information and Knowledge Management*. ACM, 175–184.

[7] Aleksandr Chuklin, Pavel Serdyukov, and Maarten de Rijke. 2013. Click model-based information retrieval metrics. In *SIGIR '13: 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 493–502.

[8] Patricia Cohen, Stephen G West, and Leona S Aiken. 2014. *Applied multiple regression/correlation analysis for the behavioral sciences*. Psychology Press.

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

[10] Marco Ferrante, Nicola Ferro, and Maria Maistro. 2014. Injecting user models and time into precision via Markov chains. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 597–606.

[11] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, and Shipeng Li. 2011. The role of attractiveness in web image search. In *19th ACM International Conference on Multimedia*. ACM, 63–72.

[12] Abby Goodrum and Amanda Spink. 1999. Visual information seeking: A study of image queries on the World Wide Web. In *ASIST Annual Meeting*, Vol. 36. 665–74.

[13] Scott B Huffman and Michael Hochster. 2007. How well does result relevance predict session satisfaction?. In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 567–574.

[14] Vidit Jain and Manik Varma. 2011. Learning to re-rank: query-dependent image re-ranking using click data. In *20th International Conference on World Wide Web*. ACM, 277–286.

[15] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[16] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1–2 (2009), 1–224.

[17] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2176–2184.

[18] Yiqun Liu, Zeyang Liu, Ke Zhou, Meng Wang, Huanbo Luan, Chao Wang, Min Zhang, and Shaoping Ma. 2016. Predicting search user examination with visual saliency. In *39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 619–628.

[19] Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Evaluating mobile search with height-biased gain. In *40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 435–444.

[20] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *22nd ACM international conference on Information & Knowledge Management*. ACM, 659–668.

[21] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), Article 2.

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

[23] Harrie Oosterhuis and Maarten de Rijke. 2018. Ranking for relevance and display preferences in complex presentation layouts. In *41st international ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 845–854.

[24] Mark Sanderson. 2010. Performance measures used in image information retrieval. In *ImageCLEF*. Springer, 81–94.

[25] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 247–375.

[26] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 555–562.

[27] Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. 2013. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *22nd International Conference on World Wide Web*. ACM, 1201–1212.

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

[29] Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 11–18.

[30] Alfan Farizki Wicaksono and Alistair Moffat. 2018. Empirical evidence for search effectiveness models. In *27th ACM International Conference on Information and Knowledge Management*. ACM, 1571–1574.

[31] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The influence of image search intents on user behavior and satisfaction. In *12th ACM International Conference on Web Search and Data Mining*. ACM.

[32] Zhijing Wu, Xiaohui Xie, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. A study of user image search behavior based on log analysis. In *China Conference on Information Retrieval*. Springer, 69–80.

[33] Xiaohui Xie, Yiqun Liu, Maarten de Rijke, Jiyin He, Min Zhang, and Shaoping Ma. 2018. Why people search for images using web search engines. In *11th ACM International Conference on Web Search and Data Mining*. ACM, 655–663.

[34] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating examination behavior of image search users. In *40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 275–284.

[35] Xiaohui Xie, Jiaxin Mao, Maarten de Rijke, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2018. Constructing an interaction behavior model for web image search. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 425–434.

[36] Yuko Yotsumoto, Takeo Watanabe, and Yuka Sasaki. 2008. Different dynamics of performance and brain activation in the time course of perceptual learning. *Neuron* 57, 6 (2008), 827–833.

[37] Jun Yu, Dacheng Tao, Meng Wang, and Yong Rui. 2015. Learning to rank using user clicks and visual features for image retrieval. *IEEE Transactions on Cybernetics* 45, 4 (2015), 767–779.

[38] Fan Zhang, Yiqun Liu, Xin Li, Min Zhang, Yinghui Xu, and Shaoping Ma. 2017. Evaluating web search with a bejeweled player model. In *40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 425–434.

[39] Fan Zhang, Ke Zhou, Yunqiu Shao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. How well do offline and online evaluation metrics measure user satisfaction in web image search?. In *41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 615–624.