Evaluating Mobile Search with Height-Biased Gain

Cheng Luo
luochengleo@gmail.com
DCST, Tsinghua University
Beijing, China

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

Tetsuya Sakai
tetsuyasakai@acm.org
Waseda University
Tokyo, Japan

Fan Zhang
franky94@gmail.com
DCST, Tsinghua University
Beijing, China

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

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

ABSTRACT

Mobile search engine result pages (SERPs) are becoming highly visual and heterogenous. Unlike the traditional ten-blue-link SERPs for desktop search, different verticals and cards occupy different amounts of space within the small screen. Hence, traditional retrieval measures that regard the SERP as a ranked list of homogeneous items are not adequate for evaluating the overall quality of mobile SERPs. Specifically, we address the following new problems in mobile search evaluation: (1) Different retrieved items have different heights within the scrollable SERP, unlike a ten-blue-link SERP in which results have similar heights with each other. Therefore, the traditional rank-based decaying functions are not adequate for mobile search metrics. (2) For some types of verticals and cards, the information that the user seeks is already embedded in the snippet, which makes clicking on those items to access the landing page unnecessary. (3) For some results with complex sub-components (and usually a large height), the total gain of the results cannot be obtained if users only read part of their contents. The benefit brought by the result is affected by user's reading behavior and the internal gain distribution (over the height) should be modeled to get a more accurate estimation. To tackle these problems, we conduct a lab-based user study to construct suitable user behavior model for mobile search evaluation. From the results, we find that the geometric heights of user's browsing trails can be adopted as a good signal of user effort. Based on these findings, we propose a new evaluation metric, Height-Biased Gain, which is calculated by summing up the product of gain distribution and discount factors that are both modeled in terms of result height. To evaluate the effectiveness of the proposed metric, we compare the agreement of evaluation metrics with side-by-side user preferences on a test collection composed of four mobile search engines. Experimental results show that HBG agrees with user preferences 85.33% of the time, which is better than all existing metrics.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; *Retrieval on mobile devices*; *Environment-specific retrieval*;

KEYWORDS

evaluation metric; mobile search; user behavior

ACM Reference format:
Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Evaluating Mobile Search with Height-Biased Gain. In *Proceedings of SIGIR '17, Shinjuku, Tokyo, Japan, August 07-11, 2017,* 10 pages.
https://doi.org/http://dx.doi.org/10.1145/3077136.3080795

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

ACM ISBN 978-1-4503-5022-8/17/08…$15.00
https://doi.org/http://dx.doi.org/10.1145/3077136.3080795

1 INTRODUCTION

With the rapid growth of mobile device usage, search engine users are able to take action to address their information needs almost anytime and anywhere, particularly with smartphones. The massive shift in search engine consumers' behavior [1] forces both industry and academia to redesign existing technologies in the context of mobile search.

Search evaluation sits at the center of Information Retrieval (IR) researches. It helps researchers measure the quality of search results and the search users' experiences. Compared to *user-oriented* evaluation methods which directly model key aspects of users' interaction process, we believe that *system-oriented* evaluation is still a necessity due to its simplicity and repeatability. In the Cranfield-style evaluation paradigm [8], once the relevance judgment is done, it can be easily reused. The evaluation metrics can be calculated with little effort and time. While user-oriented evaluation based on user study or online A/B test may take a few days or weeks, system-oriented evaluation might be more appropriate for iterative development of search engines.

Mobile search is different from desktop search in several aspects: First, the distribution of query categories is different [3, 14, 34]. Second, compared with desktop screens, mobile devices present much less content at once due to the limited screen size. Thus mobile users have to incur a higher interaction cost in order to access the same amount of information. On the other hand, modern mobile devices are usually equipped with mobile touch interactions (MTIs) [9], which provides opportunities to model users' rich interactions during search.

For both mobile and desktop search, today's search engines return far richer results than a *ten-blue-link* SERP. The results may be retrieved from various vertical information resources including news, shopping, images, knowledge cards, etc. and then federated together with traditional organic results to serve the search users. For example, Figure 1 presents the search engine result page of query *"fantastic beasts and where to find them"* (a 2016 movie) from a mobile search engine, we can see the first result of the SERP looks like a card and it includes the basic profile of the film, cast, introduction, reviews from online communities and etc. The snippet of this result is quite different from a traditional textual result, which

usually contains a title, a URL and a document summary of one or two sentences. Compared with a traditional result, it occupies a much large area. The vertical results with rich information would help the user find relevant information on SERPs with minimal effort. Modern search engines often provide "direct answers" in response to head queries, so that the user will not have to click on it to access the landing page.

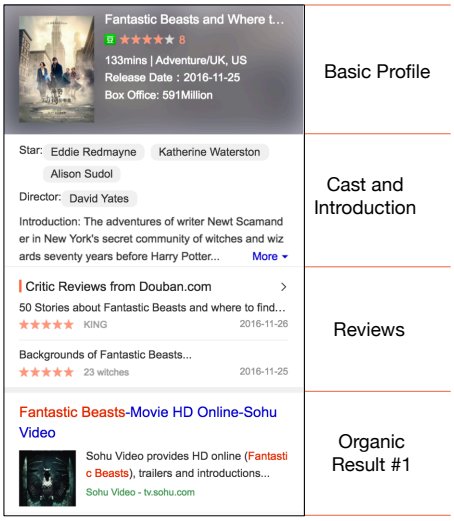


Figure 1: An example of SERP including multiple vertical results (Query: *"fantastic beasts and where to find them"*).

Due to these differences brought by "mobile v.s. desktop" and "ten-blue-links v.s. heterogenous results", traditional system-oriented evaluation method is facing a few serious challenges on mobile platforms:

Challenge 1 In mobile search, the results are heterogeneous in nature and presented in different styles. The results may occupy different percentages of the screen size, which probably means that the user's effort in examining the results varies accordingly. This poses a direct challenge to traditional rank-based evaluation measures, since the basic assumption behind them is that the ranked items and the effort spent on inspecting them are homogeneous. For mobile SERPs that include visual and heterogeneous contents, considering the height of each ranked item may be necessary.

Challenge 2 Compared with desktop search, verticals and cards are more frequently federated into mobile SERPs. In many circumstances, these results contain enough information for users' information needs. Therefore, there is no need to visit the corresponding landing pages and the effort to access useful information is reduced. However, most existing metrics do not take this factor into consideration.

Challenge 3 The internal structures of many vertical and card results are much more complex than traditional ten-blue-link results. The information contained in these complex results is also distributed among different sub-components. Therefore, it is not reasonable to assume that a user will obtain all the gain brought by the complex result once she reads one or a few sub-components. However, most existing metrics do not take the distribution of benefit/cost within results into consideration.

In the present study, we tackle the above three challenges by proposing a new evaluation metric for mobile search environment, named Height-Biased Gain (HBG). Basically, HBG follows the framework similar to Normalized Discounted Cumulative Gain (nDCG) [12], Rank Biased Precision (RBP) [24], Expected Reciprocal Rank (ERR) [6] and Q-measure [27]. It can be expressed as cumulated gain over the browsing trail of a ranked list.

With HBG, we try to address the problem of mobile search evaluation by making the following contributions:

(1) We propose to use the height of a *user browsing trail*, to estimate users' effort when they are involved in mobile search. The geometric height naturally takes both textual and non-textual contents into consideration and is therefore more appropriate to deal with heterogeneous results in mobile search (**Challenge 1**).

(2) We propose to consider a new variable, *Click Necessity* to model the cases in which users do not need to visit landing pages to derive relevant information (**Challenge 2**). We show with a labeling experiment that Click Necessity can be easily annotated with minimal effort. It can then be incorporated into HBG to improve mobile search evaluation.

(3) We investigate the internal gain distribution within complex search results and find that users' reading behavior in these results can be modeled with an inverse Gaussian distribution. It is then possible for HBG to model the benefit/cost within complex results (**Challenge 3**).

The remainder of this paper is organized as follows. We introduce our proposed metric, HBG, in Section 2. We calibrate a series of parameters in our model by conducting a small scale of user study in Section 3. Then we compare the performance of HBG and traditional rank-based metrics in predicting user performance on a test collection in Section 4. Before concluding this paper, we review related studies in Section 5.

2 HEIGHT-BIASED GAIN

2.1 A generic framework of metrics

A number of traditional evaluation metrics such as nDCG [12], RBG [24], Q-measure [27], TBG [33], U-measure [30] etc. can be expressed in a generic framework [4, 40] as:

$$\frac{1}{\mathcal{N}} \sum_{k=1}^{\infty} g_k d_k \tag{1}$$

where $\mathcal{N}$ is a normalization factor. g_k denotes the utility that the user derives from the k-th result and it is usually calculated by mapping the binary/multi-graded relevance assessment to a numeric value. d_k indicates the discount factor of the k-th result, which is often viewed as the probability that a user scans down a ranked list, growing less interested, and stops at rank k.

Our proposed metric basically follows this framework. As suggested by Carterette [4], model-based measures are actually composed from the following three underlying models:

(1) *Browsing Model*, which describes how a user interacts with a ranked result list;
(2) *Document Utility Model*, describing how a user derives utility from individual relevant documents;
(3) *Utility Accumulation Model*, which describes how a user accumulates utility in the course of browsing.

In the following parts of this section, we introduce our proposed metric, Height-Biased Gain (HBG), by describing the above three underlying models.

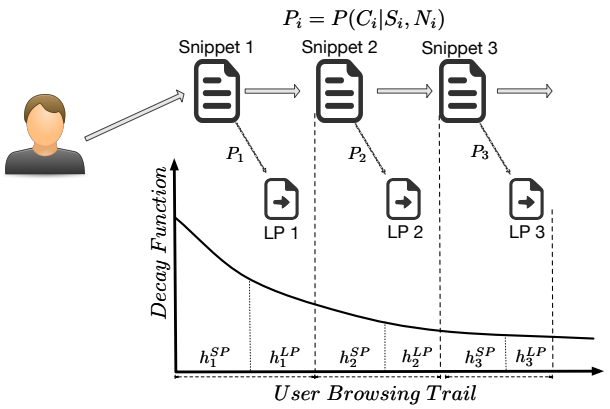


Figure 2: An illustration of user model: a user works her way down, and the value of relevant information decays with the cumulated height of viewed contents.

2.2 Browsing Model

In this study, we adopt a simple browsing model: as shown in Figure 2, the user starts at the first result of the ranking list and works her way down. For each result, the user first examines the snippet on SERP, which occupies a height of h_i^{SP} on her screen. Based on the *Snippet Relevance* (S_i) and *Click Necessity* (N_i) of the result (see Section 2.2.1), the user clicks (C_i) the hyperlink with a probability $P(C_i|S_i, N_i)$ to get more information. Otherwise, she can directly continue to examine the snippet of the next result. For the visit on the landing page of the i-th result, the user may view only part of, or the entire landing page, which is as high as h_i^{LP} on the user's screen. Eventually, the user will stop when she feels satisfied or exhausted.

2.2.1 Click Necessity. In previous evaluation metrics such as TBG [33], EBU [39], U-measure [30] and ERR [6], the probability of click is assumed to only depend on the relevance of document (R_i). We argue that for a heterogeneous result, the snippet contains much more information than a traditional textual summary. The probability of click may also be related to the presentation of a result. For example, Figure 3 presents several results of a query *"iphone 7 plus specs"* from a commercial search engine on both mobile device and desktop. Consider the first and the second result on mobile screen: the first result presents much more detailed information to satisfy the user directly. However, it is more likely that users will not click the first result's URL as much as the second result's.

To solve the problem in **Challenge 2**, we formally introduce a new variable, *Click Necessity*, to model the impact of result presentation on clicks. *Click Necessity* is defined as follows:

Definition 2.1. Click Necessity means, given the presentation of a retrieved item, how necessary it is to click it and visit the landing page to collect relevant information.

Document Relevance and Click Necessity actually describe different perspectives on a certain result: *Document Relevance* puts the emphasis on how appropriate the content of the result is for the query, while *Click Necessity* addresses the ability that the presentation of the result could fulfill users' information needs within the SERP.

With *Click Necessity*, we model the probability of click as $P(C_i|S_i, N_i)$. Following the Yilmaz et al.'s approach [39] we replace the relevance of snippet (S_i) with the relevance of document (R_i), because previous work suggests that the probability of clicks is highly correlated

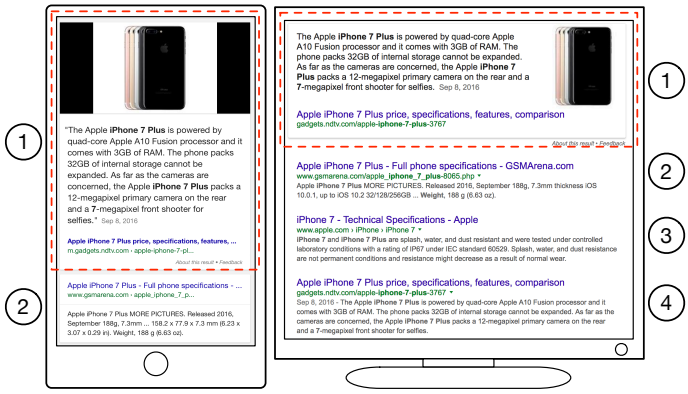


Figure 3: Results of different click necessities on mobile device and desktop (Query: "iphone 7 plus specs")

with the relevance of corresponding document.

$$P(C_i|S_i, N_i) \approx P(C_i|R_i, N_i) \quad (2)$$

In our experiment, we collected multi-level click necessity judgments following similar settings of relevance judgment in TREC [36]. Experimental results demonstrate that the click necessity could be assessed offline with relatively high inner-consistency. Details will follow in Section 3.3.

2.2.2 User Browsing Trail. Recall that our model implicitly assumes that a user scans the ranked list from top to bottom until she stops at a certain position [5]. This assumption, named *linear traversal*, forms the basis of virtually most existing evaluation metrics. Unlike rank-based evaluation metrics, TBG pays attention to the fact that the time spent at each rank differs, depending especially on document length and users' reading speed [33]. U-measure proposed by Sakai and Dou [30] used the concept of *trailtext* as a simple concatenation of the text read by users, for example, ``Snippet1 Ad2 Snippet3 Fulltext3''. This measure does not rely on the linear traversal assumption.

Inspired by the idea of *trailtext*, we introduce *user browsing trail* in our framework. Suppose that we have observed that the user has viewed a series of contents in a search task. We can then define the user browsing trail as a simple concatenation of these contents. For example, if the user examined the first snippet, the second snippet and then visited the landing page of the second result, her browsing trail could be organized as ``Snippet1 Snippet2 LandingPage2''. The height of each item in this trail, as shown in the mobile screen, can then be accumulated.

We use the geometric height of each viewed item as the basis for designing our evaluation measure based on the following considerations: First, although modern websites provide automatic screen adjustments for mobile users, reading a lot of contents on mobile screen is still difficult. If a user wants to read a document, she has to move the contents to the center of her viewport by pressing navigation buttons or scrolling on screen with fingers. This kind of "locate and read" would be done over and over again since the visible area on screen is quite small. In Figure 3, on mobile device there are only two results and the first one occupies about 3/4 of the viewport. The desktop actually presents the same list of results but the viewport accommodates four results and the first one only occupies about 1/3 of the screen area. Hence, geometric height may be a better indicator of the user's effort on mobile devices than the rank. Second, the trailtext of U-measure can only handle textual information. To accommodate the evaluation of SERPs that contain

highly visual and other non-texual (as well as textual) contents, the height of each item seems a natural choice for replacing the U-measure's "number of characters/words read." Third, compared with time in TBG, heights of contents are not user specific and can be estimated offline. This makes HBG more appropriate for system-oriented evaluation.

2.2.3 Discount Factor. In the generic framework of evaluation metrics (Equation 1), the discount factor d_k could be viewed as the probability that the user stops at a specific rank.

In traditional evaluation metrics, some discount functions depend dynamically on relevance of previously viewed documents (ERR [6] and the Sakai/Robertson *Rank-Biased Normalized Cumulative Utility* measure [31]): whenever a relevant document is found, the value of another relevant document found later is discounted. This property is recognized as *diminishing return* [5], which is in contrast with some other metrics such as nDCG and Q-measure that adopt a static discount factor based on the rank.

TBG [33] is an evaluation metric based on time rather than ranks. The discount factor in TBG is modeled as the probability that the user continues until a specific time. In M-measure [15] and U-measure [30], the discount factor is a linear function based on the length of trailtext and the offset within trailtext.

In this study, we assume that the value of a relevant information unit decays with the cumulated height of the contents the user has viewed. Hence we propose a discount function expressed in terms of *height*, $D(h)$, where h indicates the offset in *user browsing trail*. $D(h)$ could be viewed as the probability that the user continues until the cumulated height of her viewed content is as high as h. Thus, $D(0) = 1$ and when $h \to \infty$, $D(h)$ decreases to 0 monotonically.

$D(h)$ takes the difference in results' heights into consideration (**Challenge 1**). If the height of a snippet become larger, its height in the user browsing trail would probably grow, then the value of the information after this result would be discounted more. This is in line with our intuition, because it is more likely that a user would get bored or tired after reading a longer snippet, compared to a shorter one. We will provide estimations of $D(h)$ based on the the users' behavior in our user study (see Section 3).

2.3 Document Utility Model of HBG

In previous rank-based metrics, the utility (gain) of a document is usually treated as an atomic unit. That is to say, the assumption behind the evaluation metrics is that the user will get all the utility of the document once she reaches it.

While the general definition of TBG accumulates gain over time, the actual instantiation of it considered by Smucker and Clarke [33] takes the following form:

$$\frac{1}{\mathcal{N}} \sum_{k=1}^{\infty} g_k D(T(k)) \tag{3}$$

where g_k denotes the gain of the k-th result and $T(k)$ is modeled as the time to reach the result. That is, TBG assumes that, whenever the user accesses the k-th result, he acquires the gain of the entire result. However, as was discussed in **Challenge 3**, this is too strong an assumption especially for mobile devices with a small screen, in which heterogeneous contents may be presented. It is possible that the user actually obtains information from some parts of the result. This was the motivation for us to consider an internal gain distribution over the snippet and landing page.

In our framework, we assume that the utility of a certain result is distributed over its user browsing trail, including both the snippet and the landing page. The distribution is denoted as $G_k(h)$, then

$$\int_0^{h_k} G_k(h)\, dh = g_k \tag{4}$$

where g_k is the gain value of the k-th result, h_k denotes the height of the k-th results within the user browsing trail.

In practical application, h_k could be estimated with the heights of the snippet and landing page. The overall utility value, g_k, could be presented as a relevance score [4]. Our proposed Document Utility Model is able to handle different assumptions of utility distribution.

2.4 Utility Accumulation Model of HBG

Based on our Browsing Model and Document Utility Model, we introduce the proposed evaluation metric, Height-Biased Gain (HBG), by cumulating the utility over a user browsing trail.

Consider a user working her way down, with both gain and discount factors expressed in terms of height. An equivalent of Equation 1 could be expressed as:

$$\frac{1}{\mathcal{N}} \int_0^{\infty} \frac{dG(h)}{dh} * D(h) dh \tag{5}$$

where $G(h)$ denotes the cumulative gain experienced by the user at height h. Note that we accumulate the gains over heights, in contrast to TBG which accumulates them over time [33].

If we look at the gain collected on each result, let dg_k be the discounted gain of the k-th result, the metric could be written as:

$$\frac{1}{\mathcal{N}} \sum_{k=1}^{\infty} dg_k \tag{6}$$

Here, the discounted gain at rank k is given by:

$$dg_k = \int_{start_k}^{start_k + h_k} G_k(h - start_k) * D(h) dh \tag{7}$$

where $start_k$ and h_k denotes the offset of the result in user browsing trail and the height of the k-th result respectively.

3 CALIBRATION

3.1 Estimation of user browsing trail

The central idea of HBG is to cumulate document utility over a user browsing trail, where both the gain and the discount factor are expressed in terms of height.

Our Browsing Model assumes that a user works down a ranked result list in order. The *user browsing trail* is a concatenation of all the content she has viewed. If we have observations about the user's behavior, for example, eye movement recorded by eyetrackers, or mobile touch interactions (MTIs) collected on SERP, the user's browsing trail could be easily constructed by summing up the segments viewed by user. An alternative method is inferring user browsing trail based on querylogs from online users.

However, in practice, user behavior information may be unavailable. Therefore, we introduce a method to estimate the height of the user browsing trail.

Consider a user examine the k-th result on a SERP. The *expected viewed height* (evh_k) of this result in user browsing trail could be represented as:

$$evh_k = f^{SP}\left(h_k^{SP}\right) + P(C_k|R_k, N_k) * f^{LP}\left(h_k^{LP}\right) \tag{8}$$

where h_k^{SP} and h_k^{LP} denotes the height of the snippet and landing page respectively. $f^{SP}(h)$ and $f^{LP}(h)$ are the examining models of

the user on the snippet and landing page, which enable us to handle different user behavior assumptions. For example, if we assume that the user would only examine at most the first viewport of the landing page, the browsing model on Landing Page is

$$f^{LP}\left(h_k^{LP}\right) = \min\left(H_{viewport}, h_k^{LP}\right) \tag{9}$$

where $H_{viewport}$ is the height of the viewport.

The evh_k could be interpreted as the sum of two parts: we assume that the user would first examine the snippet on SERP, the expected viewed height $evh_k^{SP} = f^{SP}\left(h_k^{SP}\right)$. Then she may visit the landing page with a probability $P(C_k|R_k, N_k)$ and the expected viewed height on landing page is $evh_k^{LP} = P(C_k|R_k, N_k) * f^{LP}\left(h_k^{LP}\right)$.

Based on evh_k, the offset of each result in user browsing trail (see Section 2.4) could be estimated by summing up the expected viewed heights of the previous $k-1$ results.

$$start_k = \sum_{i=1}^{k-1} evh_i \tag{10}$$

Following previous work [33], we adopt an idealized user model representing an average user and provide a method to estimate the height of users' browsing trail. Going beyond the traditional assumption that the probability of click depends on the relevance of the document, we want to investigate whether and how the click necessity influences click behavior together with document relevance.

While traditional discount functions are based on rank or the time to reach the result, our proposed decaying function, $D(h)$, is actually defined as the probability a user continues when her browsing trail is as high as h. We need to estimate $D(h)$ based on actual user behavior.

Next, we explain how we calibrate our metric based on a laboratory user study.

3.2 User Study

To investigate users' behavior on mobile devices, we designed and conducted a laboratory user study with 20 search tasks. While this approach cannot give us a lot of data in the way search engine logs can, it enables us to collect very rich user interactions. The procedure of the experiment is shown in Figure 4.

Experiment Procedure. In our user study, the participants need to perform 20 ad-hoc search tasks on a mobile device. As shown in Figure 4, (I) Before the experiment, the participants need to complete a demographic questionnaire, which investigates their age, gender, major and familiarity with both search engines and smartphones. (II) Then they are required to watch an introduction video, which would ensure that the participants receive identical instruction. In the video, we introduce the procedure of the experiment by completing a training task. Then the participants are instructed to go through a training task to get familiar with our experimental system. (III) For each formal task, the participants are first presented with the search query (III-a). We also provide a brief explanation of the topic to make sure that the participants have the same interpretations about the information need. Then they are required to search with the query using a mobile phone in our system (III-b). More specifically, they are instructed as: *"Assume you have an information need described in the explanation. Please search with this query in our system as you are using an ordinal mobile search engine."*. A pre-defined SERP will be presented to the participants and we will explain how we generate and manipulate the SERP later. After search, a post-task questionnaire will collect their perceived quality and satisfaction during searching (III-c). While no time limits are imposed, the participants usually take about two hour to complete the 20 tasks assigned to them. Finally, they are directed to an exit-questionnaire (IV), which investigates their overall experiences in this study, for example, interest, fatigue level and etc. After collecting user behavior data from all the subjects, the relevance and click necessity of all the documents which have been presented (V) are assessed by third-party judges (see Section 3.3).

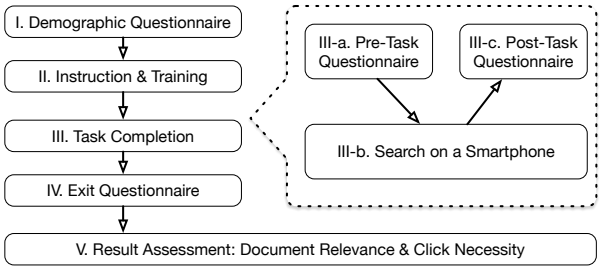


Figure 4: Experiment procedure of our user study.

Experimental System and Platform. The procedure mentioned before was conducted on an Android smart phone. It is equipped with a 5 inch screen and the resolution is 1280×720 pixels, representing one of the mainstream smartphone specifications in the year of 2016. The experiment was carried out via a Web-based system. We developed a mobile browser with Android SDK which could record the content of Webpage, MTIs (tapping, scrolling, flying and etc.) and click-throughs. All the behavior was logged by a back-end database.

Result Manipulation For each task, we have four SERPs from four major commercial search engines. Their results represent the most common search environment experienced by the majority of users. Although the presentation styles of different search engines are not exactly the same, basically the results are presented in a similar way: each result contains two parts. The first part is the a snippet on the SERP, which is intended to help the user decide if the full document should be visited or skipped. The second part is the full document itself. Due to the fact that the width of mobile devices is much smaller than that of a desktop display, the results are still ranked linearly.

To make sure that the SERPs are all adjusted for the smartphone, we crawled the search results from these four search engines by simulating a visit from the mobile device. We removed the ads, sponsored search and query suggestions to control variability and focused on the user behavior in browsing search results. All the SERPs were collected in one day of October, 2016. In this study, we focus on the user behavior on individual queries. To control variability we only present the results of the first page (about 11 results, Mean=11.48, SD=1.52) to the participants. Pagination and query reformulation was not allowed.

We then explain how we manipulate the SERPs presented to participants. The 20 tasks were divided into four groups (#1 to #5, #6 to #10 and etc.). For each participant, the queries in group was served with the SERPs from one the four search engines respectively. The user behavior data on the SERPs of 4 search engines is basically balanced, i.e. for each SERP from a certain search engine, it was presented to 11 or 12 participants. To avoid presentation order bias, the 20 tasks were rotated using a Latin Square.

Tasks and Participants The queries were sampled from query-logs of a mobile search engine. We further checked what vertical search engines these query would trigger. We found that these queries covered several major vertical result types by counting the number of tasks in each category: QA (n=10), Video (n=5), News (n=6), Image (n=6) and Knowledge (n=9). It should be noted that a task will be counted only if the SERPs from all the four search engines have presented at least one result of the corresponding vertical type. Although this taxonomy does not cover all the potential vertical types, we believe that the results belonging to these categories can represent the majority of the verticals' presentation styles. To avoid overfitting, we did not use these queries to evaluate the effectiveness of metrics (see Section 4).

We invited 43 students (20 female and 23 male, aged from 19 to 23, the median is 20) from a university via email, online forums and social networks. A variety of majors were represented across the natural science (n=13 participants), social science (n=10) and engineering (n=20). The participants reported that they were very familiar with search engine (Mean=5.68 in a 7-point Likert scale, from not familiar to very familiar) and smart phones (Mean=5.79). They were informed in advance that they would be paid $20 for the participation and all of them signed a post facto participation form revealing the content of the experiment.

3.3 Result Assessment

In our experiment, we have 80 SERPs from 4 different search engines and 918 results. On each SERP, there are 2.78 clicks on average (SD=1.70, Min=0, Max=10). In total, there are 2391 click-throughs observed.

After collecting interactions from our participants, we further assessed the results in terms of *Document Relevance* and *Click Necessity* with the help of several trained judges. The judges are graduate students whose research areas are information retrieval studies.

To make sure the assessors and the participants have similar experiences, all the assessments were done using the same type of smartphone that the participants used.

For *Document Relevance*, we used the typical four-level relevance criteria: irrelevant (R=1), marginally relevant (R=2), relevant (R=3) and highly relevant (R=4) following the TREC definition [36]. Each time only one result was shown to the assessors and the appearance of result was the same as what was shown to the participants. Assessors were required to visit the landing page before making their decisions. All the results were annotated by three assessors. The Fleiss' κ is 0.388. If there is a disagreement between judges, we use the median as the relevance of the result.

For *Click Necessity*, we adopt similar paradigm as *Document Relevance* and an independent group of assessors were hired. The difference is that we only showed the snippet on SEPR to the assessors and visiting landing page was not allowed. Then we asked them to make a decision by answering the following question with the options we provided.

Question: Assume we have an ideal document or information resource, which is presented in the same style as this result. Do you think it is necessary for the user to visit the landing page after examining the snippet?

- **Definitely Necessary (N=1)**: The snippet cannot present sufficient information and users need to visit the full document.
- **Possibly Necessary (N=2)**: The snippet could present some useful information. Some of the users may be satisfied while some others will visit the full document.
- **Not Necessary (N=3)**: The snippet is able to present sufficient information and most of the users could get enough information on the SERP. Visiting landing page is not a necessity.

Similar to *Document Relevance*, each result was assessed by three judges. The median was considered if there is a disagreement. The Fleiss' κ is 0.475, which shows that our proposed variable, *Click Necessity* could be assessed offline with a relatively high agreement.

3.4 Probability of Clicking

Based on the collected actual behavior and the assessments of results, we are able to examine how the document relevance and click necessity together impact users' clicks.

Given a SERP where the lowest clicked rank is k, we assume that the user examines all the snippets from ranks 1 through k. Thus, we are able to know all the examined snippets and clicked results in users' sessions. For every combination of Document Relevance level and Click Necessity level obtained from the judges, we computed the probability of click based on the user behavior data collected from the assessors. The probabilities are calculated by averaging #Click/#Examination over all the results. The possibilities are shown in Table 1.

Table 1: Probability of clicking given the Document Relevance (R) and Click Necessity (N).

$P(C\|R,N)$ (#Results)	N=1	N=2	N=3
R=1	0.403 (82)	0.067 (4)	0.093 (46)
R=2	0.438 (130)	0.313 (17)	0.040 (26)
R=3	0.607 (135)	0.500 (41)	0.147 (29)
R=4	0.884 (252)	0.757 (104)	0.647 (51)

We can see that the probability of clicking has a positive correlation with document relevance and a negative correlation with click necessity. An exception happens when $R = 1$, $P(C|R = 1, N = 3)$ is slightly larger than $P(C|R = 1, N = 2)$. A potential explanation is the data sparsity: in our dataset, only 4 results are labeled as $R = 1, N = 2$.

We find that even a result is highly relevant, if its click necessity score is 3, i.e. the snippet contains rich information to fulfill users' information needs, users are less likely to visit its landing page. This observation is in line with our assumption: the probability is affected by not only the relevance of the document but also the presentation on the SERP.

3.5 Modeling Browsing Behavior

Another goal of our user study is to calibrate our decay function according to the behavior of users. Our model provides no guidance regarding the form of the decay function. One of the possibilities is standard exponential decay which is used by some previous studies [10, 33, 40]. We present this decay function in terms of height h.

$$D_{Exp}(h) = e^{-h\frac{\ln 2}{half}} \tag{11}$$

where *half* is a parameter and it is usually recognized as the "half-life" of users, i.e. the height (or time in TBG) at which half of the users have stopped browsing the results.

An alternative form of decay function is provided by Luo et al. [23]. They model users' behavior in browsing hedonic contents (e.g. mobile apps for video, music, news, jokes, pictures, social net- works etc.) as a stochastic process. Although their scenario is not exactly the same as mobile search, they do have something in common: e.g. the information items are organized in a ranked sequence. The exogenous factors which would drive the user to continue or leave are also similar, for instance, content quality, visit time and etc.

Luo et al. found that the distribution over browsing length for a visit can be described by the inverse Gaussian form with high precision. We further express $D(h)$ in this form.

$$D_{IG}(h) = 1 - \Phi\left(\sqrt{\frac{\lambda}{h}}\left(\frac{h}{\mu} - 1\right)\right) - \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-\sqrt{\frac{\lambda}{h}}\left(\frac{h}{\mu} + 1\right)\right) \tag{12}$$

where Φ is the cumulative distribution function of a standard Gaussian distribution. μ denotes the mean of the original inverse Gaussian distribution, which determines the shape of $D(h)$ together with a shape parameter λ.

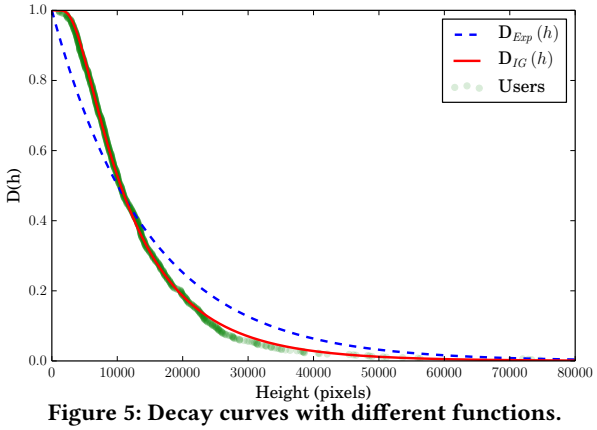


Figure 5: Decay curves with different functions.

Based on the collected MTIs, we are able to construct the browsing trail of users by concatenating all the contents they have viewed in each task. Then we fit the above two decay functions to our observations. The heights of users' browsing trail were recorded by the browser of our experimental system (see Section 3.2).

To compare the decay functions fairly, both of the decay functions were tuned to their best with the object of $L1$-norm. The optimal parameter of exponential decay is $half = 10069$ while in the inverse Gaussian decay μ and λ equals to 13510 and 23070 respectively. The decay curves of these two functions are shown in Figure 5. We test the quality of the fits by the sum of the error over all the observations. The error of exponential decay is 63.08, compared to 5.57 of inverse Gaussian decay. This means the inverse Gaussian distribution describes the probability of users' browsing behavior more accurately.

4 EXPERIMENTAL RESULTS

In this section, we first introduce the detailed parameter settings of HBG and the test collection we used. The remaining parts discuss the following two research questions:

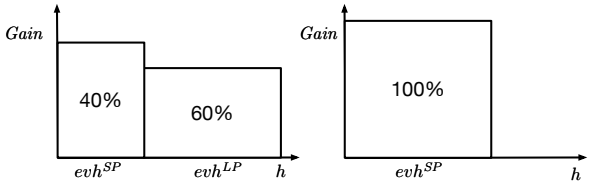
- **RQ1:** Whether and how does HBG correlate with traditional rank-based evaluation metrics in discriminating the performances of different systems?
- **RQ2:** Is HBG consistent with side-by-side user preference?

4.1 Parameter Settings of HBG

In HBG, there are several parameters and user models. Here we explain the detailed settings in our experiment.

Decay function $D(h)$: we consider the two decay functions, exponential decay ($D_{Exp}(h)$) and inverse Gaussian decay ($D_{IG}(h)$) introduced in Section 3.5, which further leads to two versions of HBG: HBG_ed and HBG_igd respectively.

Gain distribution $G_k(h)$: The benefit located at different parts in a result is difficult to measure. In this work, we only consider the difference between the snippet and the full document. We assume that the gain of the document distributes uniformly on the expected viewed height of snippet and landing page. An illustration is presented in Figure 6. For each result that has a snippet and a landing page, we assume that 40% of its gain is distributed over the snippet, while the other 60% is distributed over the landing page (Figure 6(a)). This ratio is based on Lorigo et al.'s finding that users usually spend 40% of their time looking at the snippet and the remainder elsewhere [22]. For each result that does not have a hyperlink pointing to a landing page (Figure 6(b)), we assume that all of its gain is distributed on the snippet. We can estimate more precise gain distribution by analyzing users' behavior on different types of results. We would like to explore this in our future work.



(a) $G(h)$ w/ landing page (b) $G(h)$ w/o landing page

Figure 6: An illustration of gain distribution in HBG.

Click Probability $P(C_i|R_i, N_i)$: we used the probabilities we estimated based on our user study. It should be noted that this probability may vary with search environment, enrolled ranking systems and etc. We may need to recalibrate it when the application scenario changes.

Browsing Model on SERP and landing page $f^{SP}\left(h_k^{SP}\right), f^{LP}\left(h_k^{LP}\right)$: in this study, we just use a simple browsing model $f^{SP}\left(h_k^{SP}\right) = h_k^{SP}$, $f^{LP}\left(h_k^{LP}\right) = h_k^{LP}$, indicating that users would browse the entire snippet and landing page of a result. A more fine-grained browsing model will be left for future work.

Normalization $\mathcal{N}$: Our metric is defined on the user browsing trail, which is affected by both relevance and clicked necessity. At this moment we do not see an easy way to construct an ideal result list. Therefore, following other metrics such as ERR, RBG, TBG and U-measure, we chose not to normalize our metric [29].

With these parameters and models, HBG is equipped with the ability to handle different user behavior assumptions. It is closely related to user behavior to bridge offline evaluation with real users' experiences.

4.2 Test Collection

In our study, we use SERPs from 4 mobile search engines, referred to as **A**, **B**, **C** and **D**. Our test collection was constructed by one of the four search engine companies, denoted as **A**. In the test collection, there are 50 queries. For each query, there are four mobile SERPs, among which one of them is from **A** the others belong to

the other three companies respectively. To avoid overfitting, there is no overlap between this queryset and the queries we used for the user study.

For all the SERPs, the top 5 results are assessed by the professional assessors from Search Quality Department of **A**. They adopted the same 4-level relevance criteria as TREC [36]. They also annotated side-by-side user preference between their results and the other competitors'. The assessors were given two SERPs for a given query, one from **A** and the other from **B**, **C** or **D**. The SERP pairs were presented in parallel on two smartphones and the assessors were instructed to give a confidence score according to their satisfaction with the two SERPs in a 5-point scale (−2 to +2, from much worser to much better). Each SERP pair was annotated by 7 assessors. If there is a disagreement between assessors, following Zhou et al. [41], we adopted the majority of assessors' preferences.

The click necessity of these results were assessed by the same group of judges who had annotated the results in our user study (see Section 3.3), to make sure that all the decisions are made under the same criterion.

Thus, we have 50 queries, 4 runs and their relevance and click necessity annotations. We also have 150 user preference observations between one search engine and the other three.

4.3 Correlation with Tradition Rank-based Metrics

To answer **RQ1**, here we investigate the relationship between HBG and the other ranked-based evaluation metrics.

We calculated the ranked-based metrics with the help of an open evaluation tool, `NTCIREVAL`[1]. We used several metrics available in `NTCIREVAL`, including Precision, Hit, MSnDCG [12][2], ERR [6], nERR [7], P-plus [28], Q-measure [27], AP [36], RR [35], RBP [24] and two NCU metrics [31].

For the evaluation metrics which need a cutoff, we only report the results when cutoff equals 3, since the first viewport of mobile SERP usually accommodates about 3 results, which have the largest impact on users' experiences.

In our experiment, we do not compare our metric with TBG since the instantiation of TBG reported in their paper [33] is calibrated based on a particular user study. The scenario of their user study is a traditional Web search task and its appropriateness for heterogeneous mobile search environment is unknown.

The consistency between evaluation metrics are measured by the average kendall's τ. More specifically, given two evaluation metrics, M_A and M_B, the consistency between M_A and M_B is measured by the average of kendall's τ over all the queries:

$$avg\text{-}\tau\left(M_A, M_B\right) = \frac{\sum_q \tau\left(R_{q,A}, R_{q,B}\right)}{\#queries} \tag{13}$$

where $R_{q,A}$ and $R_{q,B}$ are the rankings of different systems according to M_A and M_B respectively.

The $avg\text{-}\tau$ between all the metric pairs is presented in Figure 7, in which darker grids denote higher agreements between the corresponding metrics. It is not surprising that the metrics from the same family are more likely to be consistent with each other, for example, the nERR@3 and ERR@3 since there are only minor differences between these metrics (cutoff, decay function, normalization and etc.). The high agreement usually indicates that the metrics are sharing similar underlying user models. For example, we find that Q@3 is closely related to MSnDCG@3. The interpretation is that both Q-measure and nDCG are top-heavy metrics suitable for informational search intents, as they have been designed to consider a lot of relevant documents.

Our proposed metrics, HBG_idg and HBG_ed have a high inner-correlation (0.968) because the only difference between these two metrics is the decay function. They have a moderate correlation with the other rank-based evaluation metrics (0.637 to 0.789) indicating that HBG could get similar performance in discriminating the performance of different mobile search engines.

The correlations between HBG_igd and Hit@3 (0.637), RR (0.659) are relatively lower than other metrics. We believe that simple binary-relevance measures such as Hit and RR are clearly not adequate for our purpose.

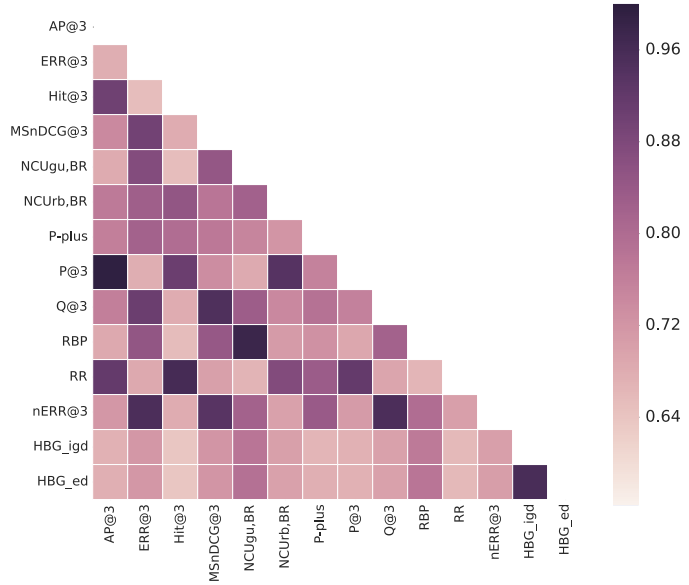


Figure 7: Consistency between evaluation metrics, which is measured by kendall's τ over 50 queries. The color indicates the correlation magnitude.

4.4 Agreement with Side-by-Side Preference

"Evaluating evaluation metrics" (or meta-evaluation of metric) is always a difficult problem in IR studies, since different evaluation metrics have different user behavior assumptions behind them. A number of methods are proposed to validate the credibility of evaluation metrics, such as Kendall's τ, Discriminative Power [25], and the Concordance (or Intuitiveness) Test [26]. These methods have been widely adopted and have aided us in gaining much insight into the effectiveness of evaluation metrics. However, they also follow certain types of user models or statistical assumptions and do not take the actual users' experiences into consideration.

Based on the user preference judgments, we look into the reliability of metrics by comparing the agreement between metrics and user preference judgments (**RQ2**). Similar approach has been adopted in a series evaluation studies to test whether the metrics line up with users' experiences [21, 32, 41]. We believe that the power to predict user preference is one of the key abilities of effectiveness measures since high agreement usually indicates that the measure is able to reflect real users' experiences.

[1] http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html
[2] MSnDCG is a Microsoft version of nDCG.

Recall that our user preference test was conducted in a 5-point scale. We first folded them into 3 classes: Worse (−2, −1), Tie (0) and Better (+1, +2). Consider the evaluation metrics, we assume if the difference between SERPs is small than a threshold δ, it should be interpreted as that there is a tie for the two SERPs in terms of result quality. More specifically, for two SERPs A and B measured by a evaluation metric m, there is a tie if $|m_A - m_B| < \delta$ when m is guaranteed to be between 0 and 1. Otherwise, we adopt $|m_A - m_B| < \delta * \max(m_A, m_B)$. In our experiment, we use $\delta = 0.05$, indicating that the difference is smaller than 5%. Then we can calculate the agreements between metrics and user preferences by counting the consistent instances. The results are presented in Table 2.

HBG_igd achieves an agreement of 85.33% while HBG_ed reaches an agreement of 80.00%, which is better than all the rank-based evaluation metrics. HBG_igd is better than HBG_ed probably due to the fact the inverse Gaussian decay function can better describe the users' leaving probabilities.

Both HBG_igd and HBG_ed outperform traditional rank-based evaluation metrics in terms of predicting user preference. We find that a potential reason is that HBG can discriminate the SERPs whose qualities are very close. For example, suppose we have two SERPs and all of their top five results are highly relevant. HBG is able to capture the impact of result presentation to user effort, while traditional rank-based metrics would probably get identical scores for the two SERPs.

We can see that among the traditional measures, Q@3 and other measures such as NCUrb,BR achieve 61-75% agreement. The next best group of traditional measures consists of NCUrb,BR, AP@3 and P@3: they achieve 47-48% agreement. The agreements for RR, RBP and Hit@3 are below 40%.

Table 2: Agreements between evaluation metrics and side-by-side user preference (Agr. and denotes #Agreements and #Disagreements respectively).

Metric	Agr.	Dis.	Rate	Metric	Agr.	Dis.	Rate
HBG_igd	128	22	85.33%	**NCUgu,BR**	100	50	66.67%
HBG_ed	120	30	80.00%	**NCUrb,BR**	92	58	61.33%
Q@3	113	37	75.33%	**AP@3**	71	79	47.33%
MSnDCG@3	111	39	74.00%	**P@3**	70	80	46.67%
nERR@3	109	41	72.67%	**RR**	58	92	38.67%
ERR@3	102	48	68.00%	**RBP**	56	94	37.33%
P-plus	101	49	67.33%	**Hit@3**	48	102	32.00%

To summarize, in this section we compare our proposed HBG and several rank-based evaluation metrics on a test collection from the following aspects: (1) We find a moderate correlation between HBG and rank-based metrics. (2) We compare the agreements between metrics and side-by-side user preference, HBG_igd achieves the highest agreement, 85.33%. (3) HBG_igd outperforms HBG_ed in terms of correlating with user preference. A possible explanation is the decay function of HBG_igd are more accurate than that of HBG_ed in describing users' reading behavior.

5 RELATED WORK

5.1 Web Search on Mobile Devices

Mobile search is different from desktop search in many aspects. According to several studies based on query logs of commercial search engine companies: Bing [34], Google [14] and Yahoo! [3, 38], the distributions of query categories on desktop/mobile devices are different. People are more likely to search for image, adult and entertainment information with mobile devices.

Compared with desktop screens, mobile devices accommodate much less content on their screens. Thus mobile users have to incur a higher interaction cost to access the same amount of information. Jones et al. found that information retrieval tasks will be harder to complete on devices with small screens [13]. Kim et al. [16, 17] conducted eye-tracking analysis of Web search users on both large and small screens. They found that with smaller screens, users exhibited less eye movement, and were slower to complete tasks.

A number of studies aim to improve users' experiences on mobile devices. Guo et al. [9] investigated mobile touch interactions (MTIs) during web search by comparing with interactions on desktop searches. They found that touch behaviors on mobile devices are significantly correlated with the document relevance. Lagun et al. [18] studied the effect of relevance in answer-like results on a mobile device. Their results indicated that users were less satisfied and tend to more time below the answer-like results. Another user study by Lagun et al. [19] put the emphasis on understanding searchers' attention with rich Ads formats on mobile devices during search sessions.

All of the aforementioned studies have aided us in gaining much insight into the users' behavior on mobile platforms and they can provide a good foundation and rationale for the construction our user behavior model.

5.2 Search Evaluation

Search evaluation is usually adopted in two ways: *System-oriented* approaches introduced a way to evaluate ranking systems with a document collection, a fixed set of queries, and relevance assessments from assessors, which is referred to as the Cranfield framework [8]. Ranking systems are evaluated with metrics, such as Precision, Recall, nDCG etc. This line of evaluation methods has the advantage that relevance annotations on query-document pairs can be reused. Beyond the Cranfield framework, IR community strives to make evaluation to correlate with real users' experiences more closely. The *user-oriented* evaluation methods, observing user behavior in their natural task procedures offer great promise in this regard. Similar evaluation protocols have also been adopted in other areas [11].

The proposed HBG is initially inspired by TBG [33] and U-measure [30]. What sets HBG apart from previous metrics is that: (1) We adopt the height of *user browsing trail* as the discount factor, instead of the integral rank in rank-based metrics, time in Time-Biased Gain (TBG) [33] and text length in U-measure [30] and M-measure [15]. (2) In HBG, we explicitly take the *Click Necessity* into consideration. The difference in click necessity may have an impact on user behavior, which further leads to various users' effort. (3) HBG takes the internal gain distribution within results into consideration while traditional rank-based evaluation metrics simply assume that users would get all the utility at a time.

Another line of research which is related to mobile search evaluation is evaluating aggregated search [41, 42]. Zhou et al. developed several metrics by extending traditional rank-based metrics (DCG, RBG, ERR and etc.) [41] and they also compared the performance of effectiveness metrics with user preferences. In the present study, we focus on evaluation in the context of mobile search, considering the impact of result presentation on users' behavior.

Recently, there have been a number of studies focusing on the "Good Abandonment" problem on both mobile and desktop searches. Given the rich presentation formats of heterogeneous results, search users may not have to click on result URLs to obtain the necessary information. A number of studies [2, 20, 37] have attempted to

detect this kind of "Good Abandonment" in both desktop and mobile search environments. However, to the best of our knowledge, we are among the first to define the variable of click necessity and take this "Good Abandonment" phenomena into consideration in search evaluation studies.

6 CONCLUSIONS AND FUTURE WORK

In summary, in this paper, we proposed a new evaluation metric, Height-Biased Gain to tackle the challenges raised in the mobile search. We proposed to use the geometric height of a user browsing trail to estimate users' effort of interactions on small screens. The height can handle both textual and non-textual contents and is able to take results with different heights into consideration. Given the snippets with rich information, we propose to consider a new variable, Click Necessity to model the cases in which users do not need to visit landing pages to obtain useful information. Also, we adopted an internal gain distribution to describe a fine-grained utility derivation course within complex search results. Based on a lab-study with 43 participants, we calibrated our metric and found that the users' reading behavior can be modeled accurately with an inverse Gaussian distribution. The effectiveness of our proposed metric was evaluated on a proprietary test collections, which contains results from 4 mobile search engines. We found that HBG could achieve a better agreement with side-by-side user preference than existing evaluation metrics.

Our study has a few limitations: (1) The parameters in our metrics are learnt from a small scale behavior dataset collected from university students. In the future, we plan to validate its appropriateness via a large scale practical log analysis. (2) In our test collection there are only 4 runs and 6 run pairs. We would like to examine the statistical properties (e.g., Discriminative Power) of HBG in the future, using more runs from a shared task.

7 ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation (61622208, 61532011, 61472206) of China and National Key Basic Research Program (2015CB358700).

REFERENCES

[1] 2015. Mobile Search Tops at Google. http://blogs.wsj.com/digits/2015/10/08/google-says-mobile-searches-surpass-those-on-pcs/. (2015). Online; Accessed: 2016-12-20.
[2] Olga Arkhipova and Lidia Grauer. Evaluating mobile web search performance by taking good abandonment into account. In *SIGIR '14*.
[3] Ricardo Baeza-Yates, Georges Dupret, and Javier Velasco. A study of mobile search queries in Japan. In *WWW '07*.
[4] Ben Carterette. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *SIGIR '11*.
[5] Olivier Chapelle, Shihao Ji, Ciya Liao, Emre Velipasaoglu, Larry Lai, and Su-Lin Wu. 2011. Intent-based diversification of web search results: metrics and algorithms. *Information Retrieval* 14, 6 (2011), 572–592.
[6] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected Reciprocal Rank for Graded Relevance. In *CIKM '09*.
[7] Charles LA Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *WSDM '11*.
[8] Cyril W Cleverdon and Michael Keen. 1966. Aslib Cranfield research project-Factors determining the performance of indexing systems; Volume 2, Test results. (1966).
[9] Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. Mining Touch Interaction Data on Mobile Devices to Predict Web Search Result Relevance. In *SIGIR '13*.
[10] Qi Guo and Yang Song. Large-Scale Analysis of Viewing Behavior: Towards Measuring Satisfaction with Mobile Proactive Systems. In *CIKM '16*.
[11] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR '16*.
[12] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002).
[13] Matt Jones, Gary Marsden, Norliza Mohd-Nasir, Kevin Boone, and George Buchanan. 1999. Improving Web interaction on small displays. *Computer Networks* 31, 11 (1999), 1129–1137.
[14] Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. Computers and Iphones and Mobile Phones, Oh My!: A Logs-based Comparison of Search Users on Different Devices. In *WWW '09*.
[15] Makoto P Kato, Virgil Pavlu, Tetsuya Sakai, Takehiro Yamamoto, and Hajime Morita. Two-layered Summaries for Mobile Search: Does the Evaluation Measure Reflect User Preferences?. In *EVIA '16*.
[16] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2015. Eye-tracking analysis of user behavior and performance in web search on large and small screens. *Journal of the Association for Information Science and Technology* 66, 3 (2015), 526–544.
[17] Jaewon Kim, Paul Thomas, Ramesh Sankaranarayana, Tom Gedeon, and Hwan-Jin Yoon. 2016. Understanding eye movements on mobile devices for better presentation of search results. *Journal of the Association for Information Science and Technology* (2016).
[18] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *SIGIR '14*.
[19] Dmitry Lagun, Donal McMahon, and Vidhya Navalpakkam. Understanding Mobile Searcher Attention with Rich Ad Formats. In *CIKM '16*.
[20] Jane Li, Scott Huffman, and Akihito Tokuda. Good abandonment in mobile and PC internet search. In *SIGIR '09*.
[21] Yiqun Liu, Ruihua Song, Min Zhang, Zhicheng Dou, Takehiro Yamamoto, Makoto P Kato, Hiroaki Ohshima, and Ke Zhou. 2014. Overview of the NTCIR-11 IMine Task. In *NTCIR '12*.
[22] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye Tracking and Online Search: Lessons Learned and Challenges Ahead. *J. Am. Soc. Inf. Sci. Technol.* 59, 7 (2008).
[23] Ping Luo, Ganbin Zhou, Jiaxi Tang, Rui Chen, Zhongjie Yu, and Qing He. Browsing Regularities in Hedonic Content Systems. In *IJCAI '16*.
[24] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* 27, 1 (2008).
[25] Tetsuya Sakai. Evaluating Evaluation Metrics Based on the Bootstrap. In *SIGIR '06*.
[26] Tetsuya Sakai. Evaluation with Informational and Navigational Intents. In *WWW '12*.
[27] Tetsuya Sakai. New Performance Metrics Based on Multigrade Relevance: Their Application to Question Answering.. In *NTCIR '04*.
[28] Tetsuya Sakai. 2007. On the properties of evaluation metrics for finding one highly relevant document. *Information and Media Technologies* 2, 4 (2007), 1163–1180.
[29] Tetsuya Sakai. 2014. *Metrics, Statistics, Tests.* Springer Berlin Heidelberg, Berlin, Heidelberg, 116–163.
[30] Tetsuya Sakai and Zhicheng Dou. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. In *SIGIR '13*.
[31] Tetsuya Sakai and Stephen Robertson. 2008. Modelling A User Population for Designing Information Retrieval Metrics.. In *EVIA '08*.
[32] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do User Preferences and Evaluation Measures Line Up?. In *SIGIR '10*.
[33] Mark D. Smucker and Charles L.A. Clarke. Time-based Calibration of Effectiveness Measures. In *SIGIR '12*.
[34] Yang Song, Hao Ma, Hongning Wang, and Kuansan Wang. Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance. In *WWW '13*.
[35] Ellen M Voorhees and others. 1999. The TREC-8 Question Answering Track Report.. In *Trec*, Vol. 99. 77–82.
[36] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval.* The MIT Press.
[37] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. Is This Your Final Answer?: Evaluating the Effect of Answers on Good Abandonment in Mobile Search. In *SIGIR '16*.
[38] Jeonghee Yi, Farzin Maghoul, and Jan Pedersen. Deciphering mobile search patterns: a study of yahoo! mobile search queries. In *WWW '08*.
[39] Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected Browsing Utility for Web Search Evaluation. In *CIKM '10*.
[40] Yuye Zhang, Laurence A. F. Park, and Alistair Moffat. 2010. Click-based evidence for decaying weight distributions in search effectiveness metrics. *Information Retrieval* 13, 1 (2010), 46–69.
[41] Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M. Jose. Evaluating Aggregated Search Pages. In *SIGIR '12*.
[42] Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M Jose. Evaluating reward and risk for vertical selection. In *CIKM '12*.