On Annotation Methodologies for Image Search Evaluation

YUNQIU SHAO, YIQUN LIU, FAN ZHANG, MIN ZHANG, and SHAOPING MA,
Tsinghua University

Image search engines differ significantly from general web search engines in the way of presenting search results. The difference leads to different interaction and examination behavior patterns, and therefore requires changes in evaluation methodologies. However, evaluation of image search still utilizes the methods for general web search. In particular, offline metrics are calculated based on coarse-fine topical relevance judgments with the assumption that users examine results in a sequential manner.

In this article, we investigate annotation methods via crowdsourcing for image search evaluation based on a lab-based user study. Using user satisfaction as the golden standard, we make several interesting findings. First, instead of item-based annotation, annotating relevance in a row-based way is more efficient without hurting performance. Second, besides topical relevance, image quality plays a crucial role when evaluating the image search results, and the importance of image quality changes with search intent. Third, compared to traditional four-level scales, the fine-grain annotation method outperforms significantly. To our best knowledge, our work is the first to systematically study how diverse factors in data annotation impact image search evaluation. Our results suggest different strategies for exploiting the crowdsourcing to get data annotated under different conditions.

CCS Concepts: • **Information systems → Evaluation of retrieval results**; *Relevance assessment*;

Additional Key Words and Phrases: Image search, offline evaluation, user satisfaction, crowdsourcing annotation

ACM Reference format:
Yunqiu Shao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2019. On Annotation Methodologies for Image Search Evaluation. *ACM Trans. Inf. Syst.* 37, 3, Article 29 (March 2019), 32 pages.
https://doi.org/10.1145/3309994

1 INTRODUCTION

With the bloom of multimedia contents on the web, image search has become increasingly important. The way to present image search results quite differs from traditional web search (see Figure 1 for an example). To be detailed, the results are placed in a 2D panel rather than a sequential list. Meanwhile, instead of document snippets, most image search engines show the snapshots along with some meta-information of images. Free from the "next page" button, image results of a new page is usually loaded just by easily scrolling down. All of these differences lead to changes both

This work was supported by Natural Science Foundation of China (grant nos. 61622208, 61732008, 61532011) and the National Key Research and Development Program of China (2018YFC0831700).
Authors' addresses: Y. Shao, Y. Liu (corresponding author), F. Zhang, M. Zhang, S. Ma, Tsinghua University; emails: shaoyunqiu14@gmail.com, yiqunliu@tsinghua.edu.cn, frankyzf94@gmail.com, z-m@tsinghua.edu.cn, msp@mail.tsinghua.edu.cn.

1046-8188/2019/03-ART29 $15.00
https://doi.org/10.1145/3309994

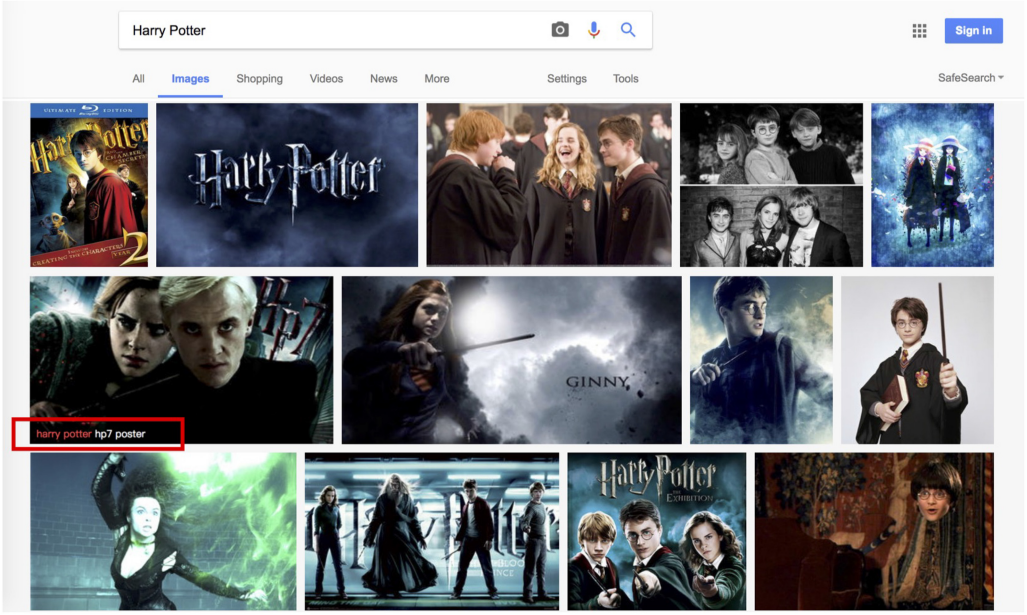


Fig. 1. An example SERP displayed by an image search engine. The red box shows the meta-information while hovering over an image result.

in the user's interaction and examination behaviors. Xie et al. [57] observe a middle-position bias in image search result pages. Since the self-contained results enable users to see and compare the image previews directly, other factors besides topical relevance, such as visual attractiveness [17, 37] and the context [50], can also affect user satisfaction of image search.

Although image search has been a very active research area in recent years, there has been little work in investigating the evaluation under the image search scenario. Both annotation protocols and evaluation metrics of image search still apply the existing standard ones developed for general web search despite those differences. Traditional metrics such as DCG [23], RBP [35], and ERR [7] assume a top-down browsing model based on query-document relevance judgments (either binary or graded). Considering the differences in the image result presentation and user behaviors, the design of both annotation and evaluation methods for image search results is an open question to be answered.

Relevance annotation is a critical part of information retrieval (IR) evaluation, since the Cranfield experiments [11]. In recent years, crowdsourcing, which is less expensive than expert efforts, has been gradually employed in the collection process of relevance judgments. Classical binary scales or ordinal scales within a small limited categories [52, 55] (usually ranging from 3 to 11) are quite common. However, the proper scale in different scenarios is still under discussion [52]. Fine-grain scales (S100) [40], which take care of more detailed perception of relevance levels, have been recently proposed in traditional (text) web search. Comparing to traditional web search results, images contain more subjective factors. The suitable relevance annotation scale is fundamental to image search evaluation yet is underinvestigated.

Satisfaction can be viewed as the golden standard in search performance evaluation. There are several works on the relationship between evaluation metrics and user satisfaction [1, 34, 43]. Mao et al. [32] point out that some traditional system-centered metrics are not well aligned with user satisfaction. In addition to traditional web search, the correlations between evaluation metrics

and user satisfaction are also investigated in mobile search [19, 26], and homogeneous and heterogeneous environments [8]. However, explicit user satisfaction has not been considered when evaluating image search results.

In this article, to shed light on the preceding research questions in image search evaluation, we mainly investigate crowdsourcing annotation methods for image results and how offline metrics perform with the annotated results, considering factors of various dimensions. We first collect explicit user satisfaction feedbacks via a laboratory user study and then employ crowdsourcing to gather the traditional four-level topical relevance judgments, quality judgments, row-based topical relevance judgments, page-based relevance judgments, and fine-grain relevance judgments. We compare the performances of offline metrics based on these different annotations by comparing their alignments to user satisfaction. To be specific, we consider the following research questions:

- **RQ1:** How do offline metrics align with user satisfaction based on the traditional four-point scaled topical relevance judgments?
- **RQ2:** How do context factors (e.g., row- or page-based relevance) influence the performance of image search evaluation?
- **RQ3:** What are the impacts of image quality for image search evaluation?
- **RQ4:** How do fine-grain relevance scales affect the performance of image search evaluation?

This article is a revised and extended edition of research that appeared at SIGIR 2018 [59]. Instead of comparing the performances of offline and online metrics, we mainly investigate the impacts of different data annotations on offline metric performances. In particular, besides four-level topical relevance and image quality judgments, we also consider factors like the context and fine-grain scales. This version extends the conference version by adding row-based, page-based, and fine-grain relevance annotations via crowdsourcing, as well as the corresponding analysis of results.

In the next section, we review related work. The details of user study and data annotation via crowdsourcing are given in Section 3. Sections 4 and 5 give experiment results and an analysis corresponding to our research questions. In Section 6, we compare different strategies for collecting data annotations via crowdsourcing. We summarize our work and directions for future work in Section 7.

2 RELATED WORK

2.1 Image Search

Image search has been shown to be a markedly active part within web search. Song et al. [48] show that queries with image intents have been second only to navigational intent queries on desktops. Further, Xie et al. [56] propose a taxonomy of image search intent, categorizing image search tasks into three groups, which are exploring, entertaining, and locating, respectively. User behaviors of image search have been studied from various dimensions. Through log analysis [4, 18, 37, 39], many interactive behavior patterns, such as query formulation, session length, hover, and click, are investigated. Park et al. [38] also do a large-scale behavior analysis based on query logs of Yahoo image search. Compared to traditional web search, the query length tends to be shorter [18, 39], and due to the image previews shown in SERPs, click becomes sparse, whereas hover becomes a quite strong signal [37, 57]. Xie et al. [57] observe a middle-position bias of user examination behavior in the image search scenario, contrary to the "Golden Triangle" phenomenon in general web search. Besides topical relevance, other factors have also been considered in image search. For example, Geng et al. [17] emphasize the importance of image attractiveness and attempt to predict the attractiveness via computational visual features and verify that the prediction results

can benefit rankings. O'Hare et al. [37] evaluate the importance of user interactive signals via normalized discounted cumulative gain (nDCG) based on a combination of relevance and image quality. But the roles of topical relevance and quality in image search evaluation have not been systematically investigated. In addition, van Leuken et al. [54] attempt to combine visual diversity to improve ranking results of image engine, and Spyromitros-Xioufis et al. [50] shed light on the context of image search result page.

Although image search has been studied from various perspectives, the evaluation of image search still utilizes standard methodologies developed for general web search without adapting to the changes we have mentioned previously. Evaluation, with no doubt, sits in the center of IR. Therefore, in our work, we focus on the evaluation process of image search, including result annotation and metric design.

2.2 Relevance

Relevance is a key notion in information science and IR in particular [44], as it is also fundamental to IR system evaluation in the Cranfield framework [11]. Typically, these judgments are made based on "topical relevance," a judgment of whether the document contains any information that is "about" the material that the "topic" is asking for. Historically, relevance judgments are made in binary scales, relevant or not. In recent years, multi-level relevance annotations have been proposed and used, but they are still coarse-grain ordinal relevance judgments. For example, three-level scales were used in TREC Terabyte Track [10], four-level scales were used by Sormunen [49], and six-level scales were used in TREC-Web Track [13]. Tang et al. [52] study participants' confidence in judgments of relevance to specific topics and find that confidence is maximized when using seven-level scales. However, Cox [14] suggests that no single number of alternatives is suitable for any situation. Besides, since the distance between the ranked categories is not well defined, mathematical operations (e.g., mean) are meaningless [45] and thus the median is used more commonly. Therefore, Maddalena et al. [31] and Turpin et al. [53] investigate the use of magnitude estimation (ME), a psychophysical scaling technique, in relevance judgments in IR and get good correlations with traditional ordinal judgments in the TREC dataset. To overcome the drawbacks of ME, for example, pre-training and detailed normalization are required, the fine-grain relevance scale (S100), ranging from 0 to 100, was proposed by Roitero et al. [40]. S100 has been verified to give annotators more flexibility than traditional coarse-grain scales but be easier for aggregation and more robust than ME. In our work, we are the first to employ the fine-grain scales to image search evaluation.

With the rapid growth of document collection size, crowdsourcing, which offers a fast, low-cost and scalable way to gather annotations, has drawn attention and gradually been used in practice in the field of IR. For example, crowdsourcing was used in the TREC Blog Track [33] and TREC Crowdsourcing Track [47]. Alonso and Mizzaro [2, 3] compare the relevance judgments collected by crowd to those made by experts assessors and claim that crowd relevance judgments can be reliable. In addition to relevance judgments, crowdsourcing is also used for evaluating interactive IR systems [60]. However, the quality of crowdsourcing annotation is always under doubt. Since cognitive bias in crowdsourcing does exist [15], proper aggregation and quality check methods should be considered. Hosseini et al. [22] propose that using expectation maximization (EM) for aggregation can outperform the majority vote (MV) method in the accuracy of relevance judgments and IR systems ranking. Kutlu et al. [25] look at the rationales to analyze disagreements and to guarantee qualities. Besides, time limits are also used for quality assurance in crowdsourcing [30]. In our work, considering the large size of image results, we utilize crowdsourcing to collect annotations, and the details will be discussed in Section 3.2.

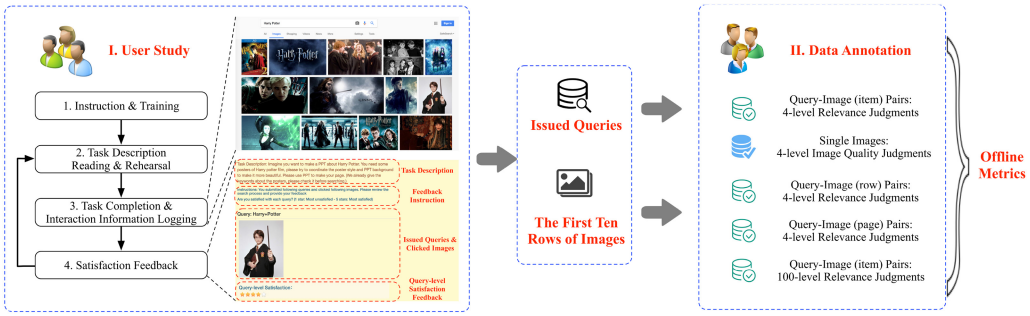


Fig. 2. Two-stage data collection procedure. The first stage is a user study, which simulates a practical image search scenario, and we collect user satisfaction feedbacks in this stage. We collect crowdsourcing annotation in stage II.

2.3 Offline Metrics

Traditional system-centric offline metrics are usually based on relevance judgments of query-document pairs from external assessors, which mainly originate from the Cranfield framework [11]. Based on binary scale relevance judgments, metrics like precision, recall, and mean average precision (MAP) are used to measure the quality of ranking algorithms. Along with the graded relevance judgments, metrics adapted to multi-levels such as nDCG [23], expected reciprocal rank (ERR) [7], and rank-biased precision (RBP) [35] have been proposed and widely used in practice. Carterette [6] develops a conceptual framework to interpret traditional offline model-based measures, mainly based on the assumption that users examine the result list in a top-down manner. Over the past decade, metrics have evolved to be gain/utility based. For example, Zhang et al. [58] propose a bejeweled player model to evaluate a web page based on a benefit-cost framework. Azzopardi et al. [5] adopt the C/W/L framework to measure search engine result pages. Besides the position in a rank list, some other aspects have been taken into consideration as the discounting factor to develop offline metrics. Time-biased gain (TBG) [46] uses time spent by the user as the basis for discounting, whereas U-Measure [42] looks at the text length. Moreover, offline metrics also change with a different search environment. Luo at el. [29] consider the height of user browsing trail, as well as click necessity, and develop height-biased gain (HBG) for the mobile search environment. Image search engines show results differently from general Web search engines, and user examination behavior also differs [57]. However, to the best of our knowledge, no offline metrics have been designed for the image search scenario up to now.

3 METHODS

In this section, we describe our data collection procedure as is shown in Figure 2. The procedure consists of two stages. In the first stage, we designed a laboratory user study, which simulated a practical image search scenario, to collect explicit user satisfaction feedback, as well as the query-image pairs. Then we exploited crowdsourcing to get data annotated from various dimensions, including topical relevance and image quality. All of our collected data are available online[1] for academic research.

3.1 User Study

We describe the details of our laboratory user study (Stage I) in this part.

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

3.1.1 Experiment Procedure. As shown in Figure 2, after reading through the experiment instructions and finishing a training task to become familiar with the study flow, each participant was required to complete 12 web image search tasks. For each task, we provided a detailed task description to give a search intent and thus simulated a practical image search scenario. The participants were asked to read the description and repeat it in their own words first to make sure that they had fully understood the task requirements. Then they would be redirected to an experimental image search system, the results of which were provided by a popular commercial image search engine.[2] The participants could submit queries, scroll up and down, click on the results, and even download the full-size images in the experimental system, just like naturally using an image search engine. Once the participants thought that the task was completed or it was difficult to find any more useful information, they could just click on the finish button to stop searching and then complete the task requirements. After that, the participants were required to provide feedback. To help them recall the search process, all of the queries and clicked images in this task were shown in the same order as they were issued or clicked. Finally, we collected five-point scaled query-level satisfaction feedback with the instructions introduced by Liu et al. [28], where 5 means the most satisfactory and 1 means the least. According to prior work, satisfaction in IR is defined as the fulfillment of a user's information need [16]. Note that for simplicity, we focus on query-level satisfaction rather than session-level satisfaction in this preliminary work.

In our user study, the experiment was conducted on a 17-inch LCD monitor with a resolution of $1{,}366 \times 768$ pixels. The search system was displayed on a Google Chrome browser, where we injected a customized JavaScript plugin into search result pages to record participants' search behaviors including scrolling, hover, click, tab switching, and mouse movement. We also recorded queries issued by the participants and some information of image in the corresponding SERPs, including the URL, the position on the result page, and meta-information returned by the system. We later downloaded all the images for data annotation (Stage II).

3.1.2 Tasks. According to the image search intent taxonomy proposed by Xie et al. [56], all the image search tasks can be categorized into three intent categories, which are defined as follows:

- **Exploring:** Users want to learn something, confirm information, or compare information by browsing images.
- **Entertaining:** Users want to relax and kill time by freely browsing the image search results.
- **Locating:** Users want to find images for further use. They already have some requirements for these images.

Following the work of Xie et al. [56], we designed 12 image search tasks (4 tasks for each category) that cover various image search intents. The tasks are demonstrated in Table 1. Note that the language we use in this user study is Chinese, so the task descriptions, search systems, and instructions are all in Chinese. We show the English translation version in this article.

Depending on different image search intents, we provide different requirements for different tasks. As shown in Table 1, for the "Exploring" tasks, the participants only need to verbally describe the information they have found or learned. In Task 1, for example, the participants are required to describe three pictures about Haikou City in words after finishing searching in this "Exploring" task. For "Entertaining" tasks, the participants could freely search and browse the images related to the topic without any further requirements. However, for "Locating" tasks, we ask the participants to make some multimedia productions, such as a slide or a poster. To guarantee that the participants only need to use images to complete their tasks, we provide a default slide or poster with some

[2] http://pic.sogou.com.

Table 1. The Tasks Adopted in Our User Study

Intent	Task ID	Task Description	Task Requirement
Exploring	1	You just received a job offer in Haikou City. You want to know more about this city (e.g., streets, landscapes, buildings).	Please describe three pictures that are impressive to you in words.
	2	You prepare to renovate a new house. You would like to compare different decoration styles (e.g., Chinese style, simple European style).	Please introduce and compare the characteristics of different decoration styles in words.
	3	You bought a white lined t-shirt yesterday, and you want to see which pants and shoes can match it.	Please describe the most frequently chosen pants and shoes style in words.
	4	You saw a beautiful flower on the way to school. The flower had a white petal and yellow stamen, and you want to find out its name.	Please find and say the name of the flower that has the characteristic described earlier.
Entertaining	5	You want to browse some posters or photos of your favorite stars.	—
	6	You want to search for some humorous pictures to relax yourself.	—
	7	You want to browse some posters or pictures of your favorite movies.	—
	8	You want to browse some pictures of your favorite cartoons.	—
Locating	9	You are a famous designer and are invited to design a poster for a dancing party that will be held this weekend. Detail requirements including dancing people and wine glasses.	Please use PPT to design your poster. (We already provide the background of the poster in PPT.)
	10	You want to write a short news report of the 2016 U.S. presidential election. Find useful pictures for your report.	Please use Word to write your news report. (We already provide the text part in Word; please find the pictures based on the text.)
	11	You want to make a PPT about Harry Potter. You need some posters of the Harry Potter film. Please try to coordinate the poster style and PPT background to make it more beautiful.	Please use PPT to make your page. (We already give the keywords about the posters.)
	12	You want to change the desktop background of this computer; the content of the background should contain the forest and blue sky.	Please try to find a high-quality picture with no watermark and change the desktop background to this picture.

Table 2. Statistics of the Dataset

Sessions (#)	Queries (#)	Images (#)
379	1,119	54,377
Query-Image (Item) Pairs (#)	Query-Image (Row) Pairs (#)	Query-Image (Page) Pairs (#)
79,337	11,190	2,238

necessary keywords and background. For instance, in Task 11, the participants are required to make a slide to introduce the "protagonists of Harry Potter." In the default slide we provide, we list the names of three characters of Harry Potter so that the participants only need to find some corresponding pictures of the characters to complete the slide.

3.1.3 Participants. Considering that students are among active image search users, we recruited 36 students (14 female and 22 male) to take part in our user study via email, online forums, and social networks. The ages of participants ranged from 18 to 25 years. Diverse majors were included across engineering, humanities, social science, and arts. All participants were native Chinese speakers, which guaranteed that they could understand the task descriptions and requirements exactly. All participants reported that they were familiar with the search engines and used Web image search engines regularly for both study and other daily purposes. Each participant was required to complete a training task and the 12 main tasks listed in Table 1. They were informed that it would take about 1-1/2 hours to complete all the tasks without actual time limits imposed, and they would be paid about $25 on the condition of completing the experiment carefully.

3.1.4 Data Cleaning. Before data annotation, we did data cleaning. We filtered out 53 search sessions because of technical problems in recording user behavior logs. Then we also filtered out the images that could not be downloaded. Table 2 shows the statistics of our dataset after filtering. Note that we focus on query-level evaluation, so more than 1,000 queries are adequate according to statistical tools [41].

3.2 Data Annotation

After collecting explicit user satisfaction feedback, as well as user behaviors in our user study, we downloaded the pictures of the first 10 rows on all SERPs shown to the experiment participants, and further hired external assessors via several popular crowdsourcing platforms in China to gather data annotations from the five various dimensions (Stage II in Figure 2). We only got images of the first 10 rows (first two pages)[3] on the SERPs annotated because the experimental search system would load only 10 rows of images for each query by default and more than 80% of images clicked by the users were from the first 10 rows according to the records in our user study.

3.2.1 Four-Level Relevance Annotation. As we mentioned earlier, the suitable number of scales is an open question [14]. For the image search scenario, few works study the scales for relevance judgment. O'Hare et al. [37] used three-point scaled relevance judgments, which are *relevant*, *moderately relevant*, and *nonrelevant*, respectively, whereas commercial image search engines have their own criteria. In our work, we utilized the following four-level topical relevance scales as shown in Table 3 with reference to the criteria of a popular commercial image search engine.

We employed Baidu Zhongbao,[4] a famous crowdsourcing platform in China, to collect four-level relevance judgments for each query-image (item) pair. It is an in-house crowdsourcing platform.

[3]In the image search engine we used, although image results are loaded by scrolling down, there is still a symbolic page number and an obvious gap between each five rows. For more details, see http://pic.sogou.com/.
[4]http://zhongbao.baidu.com.

Table 3. 4-level Relevance Scales

Score	Description
0 (Irrelevant)	The image fails to match the subject of the query (e.g., the query is "Batman" while the main object in the image is "Spider-Man").
1 (Somewhat relevant)	The image is only partially relevant to the query. Specifically, the query contains two or more objects while the image only depicts part of them (e.g., the query is "Hillary Clinton and Donald Trump debating" while the image only focuses on Trump).
2 (Fairly relevant)	Although the objects are matched between the query and the image, their modifiers are different (e.g., the query is "Red Ferrari" while the image is about "Black Ferrari". The color, the modifier, differs.)
3 (Highly relevant)	Both the objects and their modifiers in the image are perfectly matched the query.

Actually, the form of in-house crowdsourcing is quite common in China. To be detailed, the company has its own system and crowdsourced workers for several kinds of annotation tasks. We need to provide data, as well as the corresponding instructions, and communicate with them in advance to clarify the requirements (e.g., the accuracy of the annotation results). The company is responsible for the annotation process, including designing the interface, training the workers,[5] assigning HITs, quality assurance, and so on. To get annotations of four-level topical relevance, we provided the query-image (item) pairs along with detailed instructions and examples for each relevance level. To examine the accuracy, we sampled about 600 images from their annotated results and checked their correctiveness manually. The accuracy was greater than 95% and thus we accepted all of their results. We collected judgments for each pair from three different annotators and considered the median when disagreements appeared. Note that because topical relevance is query dependent, we required that the corresponding query should co-occur with the image item to be annotated on the interface.

We will further compare and discuss the details of different annotation tasks in Section 6.

3.2.2 Image Quality Annotation. Besides topical relevance, image attractiveness, in other words, image quality, has been considered in prior work [17, 37]. To further study the role of image quality in image search evaluation, we collected quality annotations of each single image in our dataset with similar criteria introduced by O'Hare et al. [37]. The instructions for image quality annotation is shown in Table 4.

Similarly, we employed Baidu Zhongbao to collect image quality judgments. We collect three judgments for each image item. Considering that the perception of quality is a bit subjective, we offered detailed instructions attached with some specific examples during the actual annotation process. Since the image quality is query independent, we preprocessed the dataset and deduplicated the images that might appear on the result pages of several different queries. In this part, the annotators could only see a single image without the query when annotating since we only focused on the quality of image itself rather than other factors like relevance.

Note that according to data analysis afterward, we found that it was difficult for an annotator to distinguish between *Professional* and *Exceptional*, so we merged this two scores into one. In other words, we obtained four-level image quality judgments (Bad, Fair, Good, Excellent) at last. It

[5]The company will also discuss the requirements with us in detail again if they encounter problems during training to control quality.

Table 4. Image Quality Scales

Score	Description
0 (Bad)	Extremely low quality, obviously watermarked, out of focus, underexposed, badly framed images
1 (Fair)	Low-quality images with some technical flaws (slightly blurred, small watermarked, slightly over-/underexposed, incorrectly framed), which are not very appealing
2 (Good)	Standard-quality images without technical flaws (subject well framed, in focus, easily recognizable, not easily perceived watermarked), low value for download or image collections
3 (Professional)	Professional-quality images (flawless framing, focus, lighting, not watermarked), which should also be somewhat attractive/appealing
4 (Exceptional)	Very appealing images, showing both outstanding professional quality (photographic and/or editing techniques) and high artistic value

Table 5. Four-Level Row-/Page-Based Relevance Scales

Score	Description
0 (Irrelevant)	The images in the row/page are totally irrelevant as a whole.
1 (Somewhat relevant)	The images in the row/page are generally related to the query terms, but the subjects are not prominent.
2 (Fairly relevant)	The images in the row/page are generally quite related to the query terms but do not fully satisfy the query requirements.
3 (Highly relevant)	The images in the row/page are highly relevant to the query as a whole and can fully satisfy the query requirements.

also makes quality judgments more comparable and combinable with four-level topical relevance judgments, which will be discussed in detail in later sections.

3.2.3 Row- and Page-Based Relevance Annotation. As Figure 1 shows, the image previews are placed in a 2D panel on SERPs. The placement enables users to easily examine and compare image results without much effort in examining the landing pages. It inspired us that the judgments made by users in the practical image search scenario can probably be affected by nearby images, such as images in the same row or images on the same page. To validate this assumption, we collected four-point scaled topical relevance judgments in each row and each page.[6] The descriptions for each scale is shown in Table 5.

Again, we hired Baidu Zhongbao to annotate the topical relevance for each row and page respectively, and we collected judgments from three workers for each. Since no previous work on the annotation of row- and page-based scores exists, the platform of Baidu Zhongbao was not able to load a row or a page automatically. Therefore, we stitched the image items of a row/page into an integral picture according to the positions we recorded in the user study. Then the workers made judgments based on these synthesized query-image pairs. Note that we choose a set-wide conjunctive definition of relevance in a row/page here, as it is not easy to compute the overall relevance from relevance scores of image items. In the next two sections, we will also compare the results of this annotation task with other row-based integration methods (e.g., Maximum, Minimum, Average), which stand for some disjunctive versions of relevance in a row.

[6] As we mentioned earlier, we have five rows in each page.

3.2.4 Fine-Grain Relevance Annotation. ME and fine-grain scales have recently been proposed for relevance judgments in IR. Prior works suggest that they are well aligned with traditional ordinal relevance scales (either binary or four level), and they give assessors more flexibility in terms of preferential judgments and enable some mathematical operations [31, 40, 53]. Roitero et al. [40] find that their fine-grain scales (S100) are more robust than ME. Inspired by their work, we applied S100 to image relevance annotation. As far as we know, we are the first to employ fine-grain relevance scales in image search, so we give the details of our method in this part.

Annotation task. Since the flexibility during relevance judgments is one of the points we expect from this experiment, unlike the detailed descriptions and corresponding examples we gave in the four-level annotation tasks presented earlier, we made instructions as simple as possible in this task. The instruction is given as follows:

- *Please move the slider to give an integer score in the range of 0 to 100 according to how relevant the image is to the query. The higher the score is, the higher relevance the image has.*

No examples are given this time. In each HIT, a query along with a list of image results (usually 10) is given. The initial score of each image is 50. The annotator needs to click on the link of an image first and then move the slider[7] on the right side of the image, which would not appear until the image is clicked, to give a score based on the image's topical relevance. Once she moves the slider to a proper position as she wants, she can click on the "CONFIRM" button under the slider. Once the score is confirmed, it cannot be modified anymore and the image would hide at the same time. The annotator can make judgments on the images of one HIT in any order she prefers. When she finishes annotating all the images in one HIT, she can click on the "SUBMIT" button at the bottom of the page to submit the results. For each valid submission of a HIT, the annotator would be paid about $0.08. In total, at least five scores were gathered for each query-image pair.

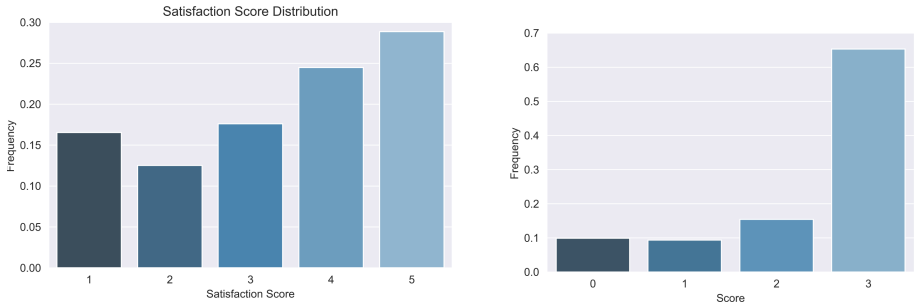
Quality assurance. To avoid potential ordering effects and first sample bias [31, 36], we used a randomized design, with images grouped into units and presented in a random order. We also included the following additional quality checks:

(1) Annotators are required to move the slider (pre-set at 50) for at least 60% of images in one HIT.
(2) The time spent in each HIT is no less than 10 seconds (usually 10 images in a unit).

If the annotator fails any of these quality checks, then no valid submission could be made and she would be assigned another HIT. We also carried some checks after collecting data, including (1) checking the submission logs of participants who completed more than 1,000 HITs (8,479 HITs overall) and (2) randomly checking 80 HITs by hand. To be specific, if the scores given by one annotator for a highly relevant image and an obviously irrelevant image of the same HIT are in a wrong order, she would fail the quality check for this HIT. If she fails more than five checks among the 80 HITs, all the annotations she has made would be abandoned. As a result, most of HITs were completed in 20 to 40 seconds, which seemed normal, and no worker failed the second manual check. Although some mistakes in a minority of HITs exist, we believe that it is a common phenomenon in crowdsourcing and their impacts on the overall results are negligible.

Crowdsourcing. Since existing in-house crowdsourcing platforms could not meet the requirements of randomized data selecting, specialized interface, and additional quality checks in this task, we utilized another open crowdsourcing platform Chinacrowds,[8] a lightweight version of

[7]We used the slider instead of the text box here to enable the worker to directly express the level of perceived relevance rather than tangle in specific numbers, and this setting is consistent with that of previous work on S100.
[8]http://www.chinacrowds.com.



(a) Satisfaction Feedbacks (b) Item-based S4 Relevance Scores

Fig. 3. (a) The distribution of satisfaction feedback in the user study. (b) The marginal distribution of item-based topical relevance scores in a four-point scale (S4).

Crowdflower[9] in China, where the designed interface and quality checks could be inserted easily. Note that we did not use platforms like Crowdflower and Amazon Mechanical Turk (MTurk)[10] because our tasks and queries were all in Chinese, whereas these platforms had few active Chinese users.

4 DISTRIBUTION

4.1 Satisfaction

We collected explicit user satisfaction feedback in the first stage, which functions as the golden standard in image search evaluation. The distribution of query-level satisfaction scores is shown in Figure 3(a). The proportions of high satisfaction scores (4 and 5) and low scores (1, 2, and 3) do not differ too much; in other words, the satisfaction distribution seems normal and balanced. And it further verifies that the task settings in our lab study are reasonable and realistic.

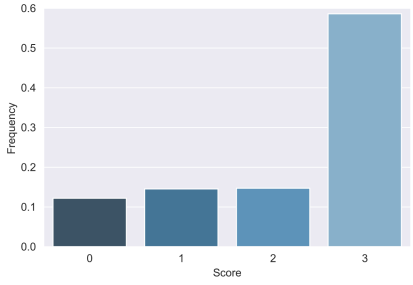
4.2 Coarse-Grain Scales

For four-level item-based relevance, row-based relevance, page-based relevance, and image quality annotations,[11] we adopt the median when there are disagreements among assessors. The Fleiss's κ of item-based relevance, row-based relevance, page-based relevance, and image quality judgments are 0.551, 0.576, 0.719, and 0.527, respectively, which all reach moderate agreements [27].

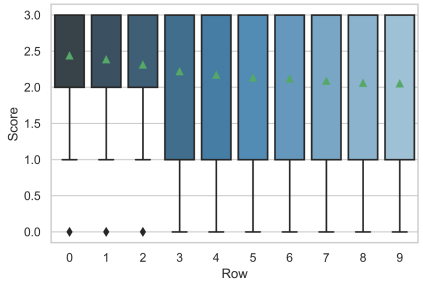
4.2.1 Item-Based Topical Relevance. Figure 3(b) gives the marginal distribution of item-based topical relevance in a four-point scale. More than 60% of images are annotated as *highly relevant*. The distribution is different from that of the traditional web search document dataset (e.g., TREC document), which has a large proportion of irrelevant documents, but it might not be surprising since the commercial image search engines are usually committed to optimizing the topical relevance of images. However, compared to the satisfaction distribution, such imbalanced distribution might weaken the discriminative power of topical relevance. Note that despite the imbalanced distribution, we did not manipulate the image results or the rankings during the user study process, because what we mainly focus on is evaluating image search results under practical search scenarios.

4.2.2 Row- and Page-Based Topical Relevance. As Figure 4(a) shows, the row-based topical relevance scores have a similar marginal distribution with item-based relevance scores in that the

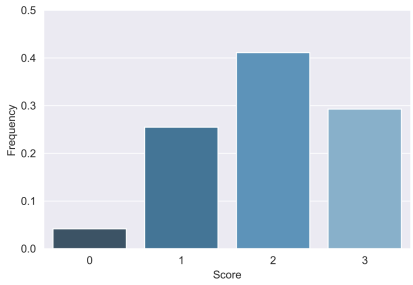
[9] https://www.figure-eight.com.
[10] https://www.mturk.com.
[11] We merge Professional and Exceptional levels when processing the results, which was explained in the previous section.



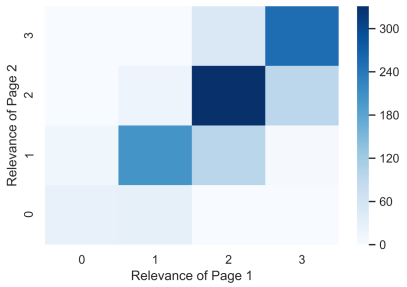
(a) Row-based Topical Relevance Scores



(b) Relevance Score Distribution of Rows



(c) Page-based Relevance Scores



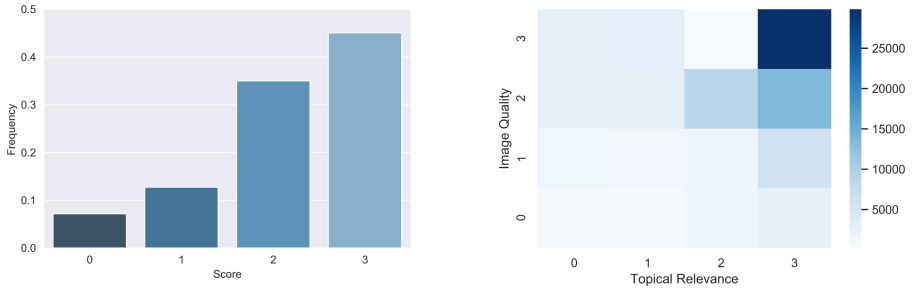
(d) Joint Relevance Score Distribution of Two Pages

Fig. 4. (a) The marginal distribution of annotated row-based topical relevance scores. (b) The distribution of relevance scores of each row; the average score is marked as triangles. (c) The marginal distribution of annotated page-based topical relevance scores. (d) The joint distribution of page-based relevance scores of page 1 and page 2. The *x*-axis represents page 1, and the *y*-axis represents page 2.

highly relevant ones account for the vast majority. However, the distribution of page-based topical relevance scores differs (see Figure 4(c)). The proportions of *somewhat relevant*, *fairly relevant*, and *highly relevant* are similar; the *fairly relevant* pages account for the most, whereas the *irrelevant* ones make up quite a small part. This indicates that most of result pages contain both irrelevant and relevant images, but relevant images are the majority. From this perspective, the distribution of page-based topical relevance scores also align with that of both item- and row-based relevance scores.

To take a further look, consider the score distribution of each row (see Figure 4(b)); we find that the first three rows have higher relevance scores. The average relevance scores decrease as the position of rows gets deeper. However, the decrease seems to converge since row 5, the first row in the second result page, which indicates that the commercial image search engine focuses on optimizing the rankings on the first page, especially the top three rows. However, the distributions of relevance scores in the first and the second pages do not show much difference. (The first page has only a slightly higher relevance score than the second page, as Figure 4(d) shows.) According to page-based relevance annotations, it suggests that the user's overall perception of the topical relevance degrees in the first two result pages are similar, because either page usually includes both relevant and irrelevant results. The exact rankings within a page seem to not matter a lot when considering the whole page.

4.2.3 Image Quality. Figure 5(a) gives the marginal distribution of image quality scores; we can observe a visible difference between item relevance and quality distributions (chi square test,



(a) Image Quality Scores (b) Joint Distribution of Relevance and Quality

Fig. 5. (a) The marginal distribution of image quality scores. (b) The joint distribution of item-based topical relevance and image quality scores. The *x*-axis represents topical relevance, and the *y*-axis represents image quality.

$p < .001$). Although image results returned by a commercial image search engine are usually of high topical relevance (score = 3), fewer images are of excellent quality (score = 3) and more images are of good quality (score = 2). To further study the relationship between topical relevance and image quality, we look at their joint distribution (Figure 5(b)), and they do not align well. Apart from the results, which are both highly relevant (score = 3) and of excellent quality (score = 3), there are many results that maintain lower quality (score < 3) despite high relevance (score = 3). We manually examine some images with a high topical relevance score (score = 3) but a lower image quality score (score < 3) and find that although most of these image contents match the related query perfectly, there are some technique flaws, such as the watermark in the images that harm the quality. The preceding observations suggest that the topical relevance and image quality are two separate facets of the image results, which may impact user satisfaction from different dimensions.

4.3 Fine-Grain Scales

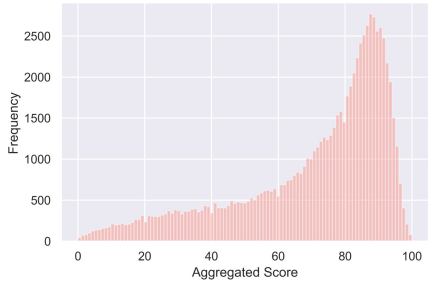
We name the fine-grain scales as S100[12] with reference to Roitero et al. [40]. Figure 6(a) shows the distribution of individual scores given by each annotator. All the scores ranging from 0 to 100 are covered, which indicates that the fine-grain scale can reflect more subtle differences in user perception of topical relevance to the query. It is interesting that there are three sharp curves at the 0, 50, and 100 points, which account for 2.9%, 1.8%, and 7.7%, respectively. There are several possible explanations for this phenomenon. By some case studies, there are some quite short and simple queries such as *"Hermione" (a character in the film Harry Potter)*, which only involve one or two items. Under such circumstances, the annotator is highly likely to give exact boundary scores, like 0 (if the image does not include the query item) and 100 (if the image content is exactly the query item). However, since our initial score for each image is 50, it requires the least effort for the annotator to give 50 points, and it might bring some bias at the beginning of annotation. We further look at whether there is a user-specific property of using the S100 scale as a ternary one (0, 50, 100) and find that there are 14 workers of which more than 50% of scores they have given are ternary. To avoid this bias, we remove all the scores they have given in all the experiments later.

As we mentioned before, we collect S100 relevance scores for each query-image pair from five different annotators. In this work, we consider the arithmetic mean[13] of scores from different

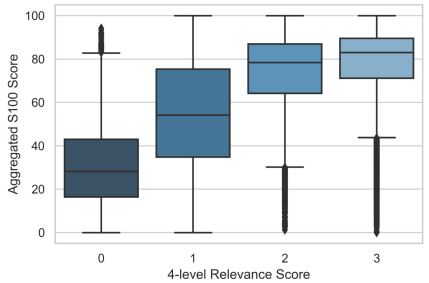
[12]The levels are actually 101, but we call it S100 for simplicity anyway.

[13]We have used several aggregation and normalization methods besides arithmetic mean, including median, min-max normalization among annotators, centralization according to the HIT and query, and other common aggregation functions,

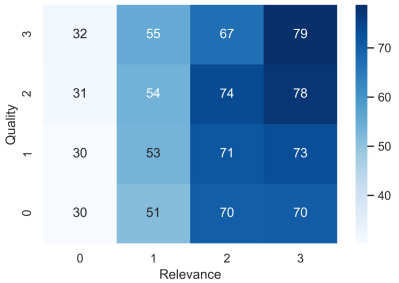
(a) Individual Relevance Score Distribution



(b) Aggregated Relevance Score Distribution



(c) Aggregated S100 Scores vs. 4-level Relevance Scores



(d) Joint Distribution of S4, S100 Topical Relevance and Image Quality

Fig. 6. (a) The distribution of individual annotated relevance scores. (b) The distribution of aggregated (average) relevance scores. The bar represents the frequency of each score. (c) Comparison of aggregated S100 scores and four-level relevance scores; the black line in the middle of each box represents the median. (d) Joint distribution of S4, S100 topical relevance and image quality; the numbers in the heat map are average S100 scores for each relevance-quality group.

annotators as the aggregated score for each query-image pair. Figure 6(b) shows the distribution of S100 scores after aggregation, and the distribution turns out to be much smoother than that of the raw scores. Most of the results are of somewhat topical relevance, and there is a peak around the score of 90, which is a rather high relevance score. It is also consistent with our observations under four-level relevance judgments. We directly compare the distribution of S100 scores compared to the four-level relevance scores (Figure 6(c)); we can observe that the S100 scores cover a larger scale of scores at each relevance level marked by four-level relevance scores, whereas the median value of S100 scores align with those relevance levels. Meanwhile, the increase of median value is nonlinear. Moreover, we are interested in the differences between two scales. Since we have annotations of image quality, which is a different dimension from topical relevance, we look at the joint distribution of image quality and topical relevance in two kinds of scales (Figure 6(d)). It is interesting that when the S4 topical relevance is low (e.g., score = 0), the S100 scores are almost the same in all quality levels, whereas when the topical relevance is high (e.g., score = 3), the S100 scores increase with image quality levels.[14] We assume that when annotating in fine-grain scales,

but it turned out that the use of arithmetic mean over raw scores performed well enough in terms of correlating with user satisfaction and was quite simple.

[14]Note that there is an exception when topical relevance = 2 and image quality = 2, which could be explained by the fact that there are few data satisfying this condition.

Table 6. Spearman's Rho (r_s) Between User Satisfaction and Metrics Calculated at "Z/S/T" Sequences Based on Four-Level Topical Relevance Annotations

	Z-Sequence	S-Sequence	T-Sequence
CG	0.180*	0.180*	0.180*
DCG@10r	0.188*	0.188*	0.190*
RBP (0.99)	**0.211***	**0.211***	**0.211***
RBP (0.8)	0.171*	0.176*	0.177*
RBP (0.5)	0.146*	0.147*	0.155*
RBP (0.1)	0.126*	0.126*	0.128*
ERR	0.122*	0.122*	0.123*
MAX	0.044	0.044	0.044
AVG	**0.193***	**0.193***	**0.193***

*Correlation is significant at the $p < 0.01$ level.

the annotators first prioritize the topical relevance factor, and other dimensions (e.g., image quality) may be further considered if certain relevance requirements are met. Thus, fine-grain scales could allow more freedom and capture annotators' more subtle perceptions.

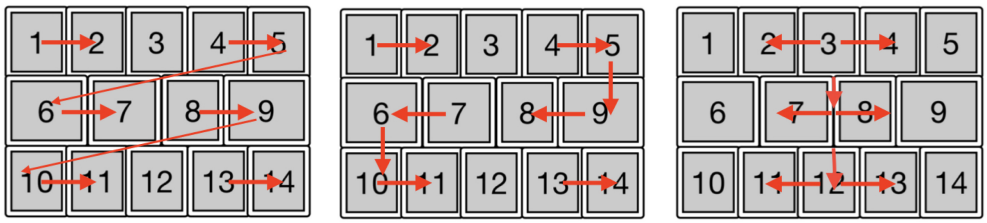
5 OFFLINE METRICS UNDER DIFFERENT JUDGMENTS

In this section, we examine how offline metrics correlate with user satisfaction (collected in Stage I) on the condition of different annotations (collected in Stage II) and attempt to answer the research questions we propose in Section 1. With user satisfaction widely considered as the golden standard in user-centric search evaluation [1, 8, 34, 43], we utilize Spearman's rank correlation coefficient to analysis how offline metrics reflect user satisfaction. In addition to Spearman's rank correlation coefficient, we have also calculated Pearson's correlation coefficient and Kendall's tau, and the overall trends are similar. Since the Spearman's rank correlation test does not carry any assumptions about the distribution of the data and is useful to analyze whether one variable is monotonically related to the other one, we use it as our primary analysis tool in the following experiments. Moreover, we also calculate the significant level of difference between correlation coefficients with reference to Cohen and Cohen [12].

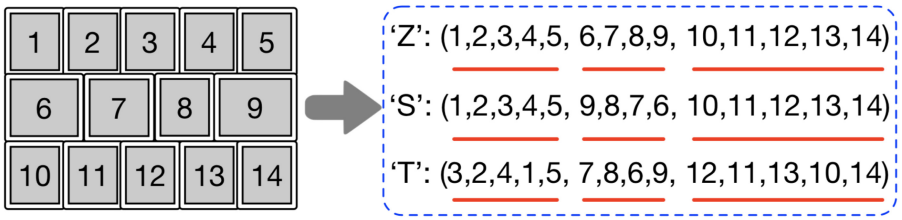
5.1 Comparison Across Offline Metrics Under Traditional S4 Topical Relevance

Based on the traditional four-level topical relevance annotations of the top 10 rows of images, we first compute several typical offline evaluation metrics that are widely used in general web search, including CG, DCG, RBP, and ERR according to the original rankings of individual image results given by the image search engine. Besides, we compute another two simple metrics—maximum (MAX) and average (AVG)—of image annotations. With reference to the previous work [8], we investigate the effects of the evaluation depth for DCG and find that DCG aligns with user satisfaction best when calculated at the top 10 rows. We use "row" rather than "rank" as a measurement of the evaluation depth here considering that the number of images varies in different rows on SERPs. For example, DCG@10r means the DCG calculated at the top 10 rows of images. Additionally, we normalize all the metrics by the number of images.

The results are shown in Table 6 (see the Z-Sequence column). We consider several typical persistency parameters p (e.g., 0.99, 0.8, 0.5, 0.1) for RBP and find that the RBP metric reaches higher correlation with user satisfaction with the increase with p. As a result, RBP (0.99) has the highest correlation coefficient among all the metrics. Meanwhile, the metric that has a slower decay rate



(a) Z-Sequence (b) S-Sequence (c) T-Sequence



(d) Results of "Z/S/T" Sequences

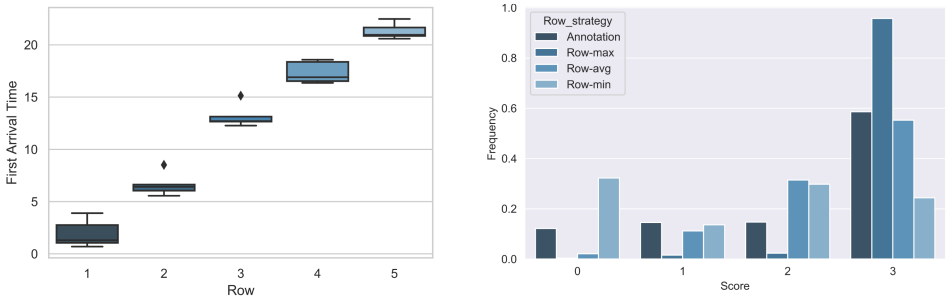
Fig. 7. (a) An example of Z-Sequence. The arrows represent how the user examines the results. (b) An example of S-Sequence. (c) An example of T-Sequence. (d) An example of how we obtain "Z/S/T" examination sequences.

(e.g., RBP (0.99), DCG, CG) shows better performance than that emphasized in the very top results (e.g., ERR). The results indicate that users tend to be patient and examine lots of images. This may be a limitation of our lab study since practical image search users become impatient sometimes. However, to collect the explicit feedbacks of satisfaction, we have to use lab-based study design. We may use some more practical experimental designs, such as field study [9, 21, 24, 51], in the future. It is also interesting that the AVG metric has high correlations among these metrics, only second to RBP. This indicates that the user might make satisfaction decisions based on the integral results (e.g., a row or a page), whereas the internal rank might not matter so much as in general web search.

Considering the middle-position bias of user examination behavior in image search [57], we compare the metric results under the assumption of three kinds of examination sequences. Figure 7 presents an example of how we obtain "Z/S/T" sequences from a two-dimensional results placement. To be specific, the "Z-Sequence" is the "original" sequence that just concatenate rows into a list (Figure 7(a)). As for "S-Sequence," it reverses the image rankings in all even lines (Figure 7(b)). As for "T-Sequence," we assume that users start to examine each row from the middle of each row and then extend to the left and right sides. They move to the next row after examining results of the current row (Figure 7(c)). The Spearman's rho (r_s) between user satisfaction and metrics based on the three sequences is shown in Table 6. Against our assumptions, there is little difference among different sequences. On the one hand, according to our analysis in Section 4, most of the images in the top rows are *highly relevant* and most of the metrics are head weighted, which makes the changes inside rows trivial. On the other hand, it indicates that users may not follow a specific sequence when examining a row.

Therefore, we only consider the original sequence, Z-Sequence, in our experiments later.

Existing metrics are mainly designed for general web search based on the sequential ranking list, without considering the two-dimensional result placement in image search. However, according to



(a) First Arrival Time of Images in the Top Rows (b) Distribution of Row-based Annotation and Integration

Fig. 8. (a) First arrival time of images in the first five rows; the black line in the middle of each box represents the median. (b) The marginal distribution of row-based annotation, Row-MAX, Row-AVG, Row-MIN topical relevance scores.

Table 7. Spearman's Rho (r_s) Between User Satisfaction and Two-Dimensional Offline Metrics Based on Four-Level Topical Relevance Annotations

	Z-Sequence	Row-MAX	Row-MIN	Row-AVG
CG	0.180*	0.180*	**0.197***	0.190*†
DCG@10r	0.188*	0.177*	**0.189***	0.184*
RBP (0.99)	0.211*	0.179*	**0.211***	0.193*
RBP (0.8)	0.171*	0.176*	**0.203***	0.186*
RBP (0.5)	0.146*	0.170*	**0.182***	0.169*
RBP (0.1)	0.126*	**0.168***	0.163*	0.140*
ERR	0.122*	**0.168***	0.167*	0.145*
MAX	0.044	0.044	**0.173***†	0.162*†
AVG	0.193*	0.181*	**0.211***†	0.193*

*Correlation is significant t the $p < 0.01$ level.
†Difference between r_s based on row integration and that based on Z-Sequence is significant at the $p < 0.05$ level.

an eye-tracking study of Xie et al. [57], we find a trend of examining images row by row by looking at the first arrival time of images on different rows, as Figure 8(a) shows. Thus, we assume that users might examine image results in a row-based method. Given this situation, we adapt these metrics to the changes in image search. Since changing rankings inside the row has not shown much influence on results (see Table 6), we use three order-independent integration methods for each row, which are the maximum (MAX), the minimum (MIN), and average (AVG) of image scores in the row. Then the two-dimensional offline metrics can be calculated as a ranking list with the integrated results for a row. The Spearman's rho between user satisfaction and these two-dimensional metrics are shown in Table 7. We also compare their correlation coefficients with those of metrics computed based on Z-Sequence, to see whether there are significant differences after integrating rows.

Row-based integration methods have different impacts on different metrics. For metrics that have no or slower decay, like CG, DCG, RBP ($p > 0.5$), and AVG, the row-based minimum integration shows the best performance, followed by the row-based average integration. Since these metrics mainly highlight gains on the result page and there are large proportion of images that are highly relevant, the image that has a lower relevance score might affect the user's perception of gain in a row and, further, satisfaction with the overall results. Meanwhile, metrics like ERR and

Table 8. Spearman's Rho (r_s) Between User Satisfaction and Metrics Calculated at Z-Sequence, Row-Based Integration Methods (We Only Show the Highest r_s of Three Methods Here), and Row- and Page-Based Annotations

	Z-Sequence	Row-ITG	Row-ANT	Page-ANT
CG	0.180*	0.197*	**0.232***†	0.228*†
DCG@10r	0.188*	0.189*	0.225*	**0.227***
RBP (0.99)	0.211*	0.211*	**0.232***	0.226*
RBP (0.5)	0.146*	0.182*	0.212*†	**0.226***†
ERR	0.122*	0.167*	0.208*†	**0.227***†
MAX	0.044	0.173*†	**0.274***†	0.214*†
AVG	0.193*	0.211*†	**0.232***†	0.228*

*Correlation is significant at the $p < 0.01$ level.
†Difference with r_s based on Z-Sequence is significant at the $p < 0.05$ level.

RBP ($p = 0.1$), which model users as less patient, are more strict to different rows, so the maximum of relevance level in a row may better reflect the perceived gain. However, most of the improvements are not significant compared to the original list (Z-Sequence). There are several possible explanations for this result. For one thing, the four-level topical relevance could not distinguish images well, considering the large proportion of highly relevant images, which further weaken the discriminative power of offline metrics. For another, we assume that these simple integration methods might not well reflect the impact of context images on user perception of relevance and could not properly represent the topical relevance of a whole row. To verify these assumptions, we then collected annotations considering different factors and compare the performance of offline metrics.

5.2 Row- and Page-Based Relevance vs. Item-Based Relevance

It is convenient for users to compare image results directly on the result page thanks to the image previews placed in a panel. Thus, perception of one image item is highly likely to be influenced by other images nearby. Instead of the single image item, a user might make decisions according to a group of images. To address RQ2, we investigate topical relevance of two typical types of groups: the row and the page. Note that the two-dimensional results of a query turn to be a sequential list based on the row- and page-based relevance. The Spearman's rank correlation coefficients between user satisfaction with metrics calculated at item-, row-, and page-based relevance are shown in Table 8. All of the metrics computed on the basis of row-based relevance seem to have better correlations with user satisfaction, and most of these improvements are significant compared to "Z-Sequence." The results also suggest that users tend to examine results in a row-based method and make decisions according to the overall relevance level rather than independent image items.

Note that row-based relevance also outperforms row-based integration in all of the metrics. We look at the marginal distribution of row-based annotation scores and the three integrated scores[15] (Figure 8(b)), and observe that there are significant differences between row-based integration results and row-based annotation scores (chi square test, $\chi^2 = 6369.2$, $p < .001$ for Row-MAX, $\chi^2 = 3176.5$, $p < .001$ for Row-AVG, $\chi^2 = 7691.7$, $p < .001$ for Row-MIN). On the condition of 4-level topical relevance, Row-Max integration tends to assign rows as *highly relevant*, which overestimates the relevance level. Row-AVG makes the relevance scores more balanced but still distributes differently from the annotated results. In particular, Row-AVG integration assigns rows of

[15]We round off the decimal numbers here.

Table 9. Spearman's Rho (r_s) Between User Satisfaction and Metrics Calculated at Topical Relevance (TR), Image Quality (IQ), and Combined Relevance (CR)

	TR	IQ	CR
CG	0.180*	0.306*†	**0.341***†
DCG@10r	0.188*	0.310*†	**0.343***†
RBP (0.99)	0.211*	0.315*†	**0.345***†
RBP (0.5)	0.146*	0.224*†	**0.256***†
ERR	0.122*	0.183*†	**0.227***†
MAX	0.044	0.010	0.057
AVG	0.193*	0.303*†	**0.335***†

*Correlation is significant t the $p < 0.01$ level.
†Difference with r_s based on TR is significant at the $p < 0.05$ level.

lower relevance levels (*irrelevant*, *somewhat relevant*) as higher scores (*fairly relevant*). The results of Row-MIN also differ a lot from the annotation results in terms of marginal distribution. But compared to the other two integration methods, it gives more weight to less relevant results and makes the distribution more balanced as a result. It explains that the Row-MIN method outperforms the other two integration methods a bit in Table 7 but is still not as good as the annotated row-based relevance scores. The results also confirm our explanation that simple integration of item-based judgments could not well reflect the context impacts and further could not well represent the integral relevance of a row.

However, the page-based relevance also brings some improvements, especially benefit metrics like ERR and RBP ($p = 0.5$). As we have mentioned before, these metrics model users as less patient and strict to the examination depth. Based on page-based relevance, images on the first two pages are considered, which is consistent with our observations that users tend to examine deep in image search. However, it does not perform as well as row-based relevance annotation for most of the metrics, which may be because there are about 30 to 50 images on a page, which are too many for a user to examine at one time. Examining results in a smaller group is more realistic. Regarding RQ2, we find that when a user examine results, the context factors like other images in the same row affect the user's perception of the image result. Users are more likely to make decisions based on row- or page-based relevance compared to the relevance of each independent image item.

5.3 Image Quality vs. Topical Relevance

Besides the context, some other aspects, such as image quality, that influence a user's measurement of general relevance may also exist, inspired by previous works [17, 37]. To address RQ3, we first compare metrics based on different judgments of Topical Relevance (TR) and Image Quality (IQ). We also employ a simple heuristic method to create a Combined Relevance (CR), which is the minimum of TR and IQ. Actually, we have tried several common heuristic methods, including product, weighted mean, maximum, minimum, and simple map with reference to prior work of O'Hare et al. [37]. As a result, metrics calculated based on the minimum of TR and IQ align with user satisfaction the best. Our intuition is that an image that is useful to users should both be highly topically relevant to the query and have high quality. The Spearman's rho between user satisfaction and metrics calculated at different judgments is shown in Table 9. The metrics based

Table 10. Spearman's Rho (r_s) Between User Satisfaction and Metrics Calculated at Topical Relevance (TR), Image Quality (IQ), and Combined Relevance (CR) in Three Search Intent Scenarios

	Exploring			Entertaining			Locating		
	TR	IQ	CR	TR	IQ	CR	TR	IQ	CR
CG	0.203*	0.129	**0.206***	0.068	**0.256*†**	0.252*†	0.036	**0.328*†**	0.308*†
DCG	0.205*	0.146*	**0.207***	0.124	**0.280*†**	0.277*†	0.043	**0.337*†**	0.320*†
RBP (0.99)	0.212*	0.152*	**0.215***	0.125	0.276*†	**0.280*†**	0.081	**0.356*†**	0.323*†
RBP (0.5)	**0.195***	0.112	0.163*	0.174*	0.267*	**0.270*†**	0.065	**0.258*†**	0.239*†
ERR	**0.156***	0.096	0.142*	0.175*	0.197*	**0.216***	0.097	0.223*†	**0.247*†**
AVG	0.210*	0.133	**0.211***	0.063	0.241*†	**0.242*†**	0.050	**0.333*†**	0.295*†

*Correlation is significant at the $p < 0.01$ level.
†Difference with r_s based on TR is significant at the $p < 0.05$ level.

on CR have much higher correlations than those based solely on TR or IQ. This suggests that both TR and IQ play important roles in the measurement of general relevance and corresponding satisfaction for image results. It is a bit surprising that metrics based on IQ perform better than those based on TR. We assume that because of the imbalanced distribution of TR of image results, TR of an image alone has poor discriminative power. Meanwhile, the image result itself contains not only information but also aesthetic value. It is reasonable that IQ matters during the process of satisfying a user's need.

Following the intent taxonomy [56], we divide our image tasks into three intent groups: Exploring, Entertaining, and Locating. In this part, we take a deep insight into the performance of metrics and the impact of TR and IQ in different search intent scenarios. In the dataset described earlier, there are 373/287/459 queries for the Exploring/Entertaining/Locating categories, respectively. Table 10 gives results. Although CR performs better overall, the performance of these three judgments are different in different search intent scenarios. For Exploring tasks, TR and CR show better performance than IQ, and CR does not bring any significant improvements. This is reasonable since users intend to learn something under this intent. Even if there are some flaws in a relevant image, users can obtain useful information from the image to fulfill their needs. In this case, TR plays a much important role than IQ. However, things change when it comes to the Locating tasks. Since the users always need to download images for some further use, they tend to be more strict to IQ. For example, in one of the Locating tasks, the participants are required to make a slide about *Harry Potter*. To make it more elegant, the participants may need to find some posters of *Harry Potter* films. Despite that an image is highly relevant to the query "Harry Potter," users may not be satisfied with the image if it has some flaws, such as a watermark. Given this reason, metrics based on TR perform far worse than those based on IQ or CR. As for Entertaining tasks, users are instructed to freely browse the image results to relax. Sometimes they just want to look through several photos of their favorite stars. In that case, topical relevance is related to whether users can find something they are interested in, which usually seems ambiguous and general in this scenario, whereas IQ is closely related to their enjoyment and furthers their satisfaction. Therefore, in the Entertaining search intent scenario, both highly relevant and highly qualified images may make users more satisfied. As a result, CR performs best at most of the time, and metrics calculated based on image quality alone correlate with user satisfaction much better than topical relevance.

Regarding RQ3, we find that IQ is a non-negligible factor when evaluating image search results, and judgments combining TR and IQ can outperform either independent one in

Table 11. Spearman's Rho (r_s) Between User Satisfaction and Metrics Based on Four-Level (S4) Relevance and S100 Relevance Annotations

	S4 Relevance	S100 Relevance
CG	0.180*	**0.334***†
DCG@10r	0.188*	**0.348***†
RBP (0.99)	0.211*	**0.342***†
RBP (0.5)	0.146*	**0.252***†
ERR	0.122*	**0.236***†
MAX	0.044	**0.299***†
AVG	0.193*	**0.326***†

*Correlation is significant at the $p < 0.01$ level.
†Difference is significant at the $p < 0.001$ level, compared to the same metric based on the S4 relevance annotations.

overall tasks. Moreover, the importance of TR and IQ will differ in different search intent scenarios.

5.4 S100 Relevance vs. S4 Relevance

Considering the imbalanced distribution of four-point scaled relevance, the coarse-grain scales might hurt the distinctive power of topical relevance. Further, we investigated fine-grain scales. To address RQ4, we compute metrics based on S4 relevance and S100 relevance annotations. Note that to avoid overflow under a [0, 100] scale, we normalize all of the scores into [0, 1] using Min-Max normalization and apply the normalized results to metrics.

Table 11 shows Spearman's rank correlation coefficients between metrics and user satisfaction. All of the metrics calculated based on S100 relevance have significantly higher correlations with user satisfaction compared to the corresponding metrics based on S4 relevance. When there are only four levels for annotators to make relevance judgments, they can hardly distinguish among images of relatively high topical relevance, which results in the imbalanced distribution (Figure 3(b)), and it also harms the discriminative power of topical relevance. However, things change when using fine-grain scales. According to Figure 6(b), scores cover the whole [0,100] scale and the aggregated score distribution seems smoother. The annotators can make further distinctions among relevant images according to their perception. Regarding RQ4, we assume that the fine-grain relevance scales can better reflect the user perception of relevance in the practical image search scenario. Thus, fine-grain relevance scales have a greater discriminative power, and metrics calculated based on them can significantly better reflect user satisfaction. In addition, it is worth mentioning that the performance of metrics based on S100 relevance scores are close to those based on combined relevance. As we have analyzed before, image quality is part of general relevance of the image result, and when annotators need to make further distinction of topical relevance, they might consider some quality factor without consciousness.

Although row-based integrated scores do not perform as well as row-based annotated scores on the four-point scale, we still consider the row-based integration as a simplified way to combine with row-based context. We compute two-dimensional metrics with three row-based integration methods based on fine-grain annotation results. Results are shown in Table 12. In general, the performance of evaluation metrics calculated at S100 relevance scores have all been improved after using row-based integration. This result reflects the impact of the row-based context. We assume that because fine-grain scales can capture more subtle differences in topical relevance and

Table 12. Spearman's Rho (r_s) Between User Satisfaction and Two-Dimensional Offline Metrics Based on S100 Relevance Annotations

	Z-Sequence	Row-MAX	Row-MIN	Row-AVG
CG	0.334*	**0.366***†	0.275*†	0.329*†
DCG@10r	0.348*	**0.361***	0.280*†	0.333*†
RBP (0.99)	0.342*	**0.365***	0.270*†	0.327*†
RBP (0.5)	0.252*	**0.329***†	0.267*	0.328*†
ERR	0.236*	0.309*†	0.269*	**0.321***†
MAX	0.299*	0.299*	0.336*	**0.352***†
AVG	0.326*	**0.365***†	0.270*†	0.327*

*Correlation is significant at the $p < 0.01$ level.
†Difference between r_s based on row integration and that based on Z-Sequence is significant at the $p < 0.05$ level.

have stronger discriminative power, the role of row-based integration has been played out. It is also interesting that the most appropriate integration method differs corresponding to the original metrics. We can see that for metrics like CG, DCG, RBP, and AVG, the most relevant image in the row determines the relevance degree of this row, whereas for metrics like ERR and MAX, the relevance degree of a row is influenced by all of the images. Since metrics such as CG, DCG, RBP (0.99), and AVG emphasize the cumulative gain of rows, the most relevant one in the row stands out and becomes representative. Meanwhile, metrics like ERR and MAX tend to be more strict to different rows; in other words, the gaps between the weights of different rows are significant and they prefer to examine fewer rows, so it becomes more cautious to measure the integral relevance degree of a row and Row-AVG integration becomes more suitable for these metrics.

Note that the proper row-based integrations methods differ a lot from that when using S4 relevance scales (see Table 7), which a little surprising. We assume that because more than 80% of rows are assigned as highly relevant when using Row-MAX integration based on S4 relevance annotations, which disables the discriminative power of TR and offline metrics, the minimum relevance score of one row becomes representative instead for metrics like CG, DCG, RBP, and AVG on consequence. However, when we get more subtle distinctions among *relevant* images, the maximum of the row becomes more representative and discriminative, and the improvements of Row-MAX integration become significant. This indicates that users still care more about positive gains during the process of examining results. However, for metrics like ERR and MAX, when the relevance score of each image contains more fine-grain factors, considering all of the images in a row rather than one item makes these metrics more stable.

Even though metrics based on S100 relevance annotations alone can achieve similar performance with those based on S4 CR, we are still interested in whether performance can be improved when combining S100 TR (TR_{S100}) and IQ. First, we experimentally compared several heuristic combination methods (e.g, maximum, minimum, product, weighted mean.) and found that the weighted mean function led to the best performance. In particular, we first normalize all the annotated scores into the scale of [0, 1] and calculate the CR (CR_{S100}) by the following formula:

$$CR_{S100} = \omega \frac{TR_{S100}}{100} + (1 - \omega)\frac{IQ}{3}$$

We do a simple grid search from 0 to 1 with a step size of 0.1 and find that when ω equals 0.6, most of the metrics have the best correlations with user satisfaction in our dataset. The weight coefficient balances the factor of topical relevance and image quality. It shows a bit more weight on

Table 13. Spearman's Rank Correlation Coefficients (r_s) Between User Satisfaction and Offline Metrics Based on S100 Topical Relevance (TR_{S100}) Alone, a Combination of Four-Level Relevance and Image Quality (CR_{S4}), and a Combination of S100 Relevance and Image Quality (CR_{S100})

	TR_{S100}	CR_{S4}	CR_{S100}
CG	0.334*	0.341*	**0.382***†
DCG@10r	0.348*	0.343*	**0.390***†
RBP (0.99)	0.342*	0.345*	**0.388***†
RBP (0.5)	0.252*	0.256*	**0.285***†
ERR	0.236*	0.227*	**0.256***
MAX	0.299*	0.057†	**0.345***†
AVG	0.326*	0.335*	**0.378***†

*Correlation is significant at the $p < 0.01$ level.
†Difference is significant at the $p < 0.05$ level, compared to the same metric based on the TR_{S100}.

Table 14. Spearman's Rank Correlation Coefficients (r_s) Between User Satisfaction and Two-Dimensional Offline Metrics Based on CR_{S100}

	Z-Sequence	Row-MAX	Row-MIN	Row-AVG
CG	0.382*	**0.404***†	0.343*†	0.380*
DCG@10r	0.390*	**0.399***	0.344*†	0.379*
RBP (0.99)	0.388*	**0.403***	0.342*†	0.379*
RBP (0.5)	0.285*	**0.364***†	0.323*	**0.364***†
ERR	0.256*	0.336*†	0.318*†	**0.353***†
MAX	0.345*	0.345*	**0.348***	0.369*
AVG	0.378*	**0.403***†	0.341*†	0.378*

*Correlation is significant at the $p < 0.01$ level.
†Difference between r_s based on row integration and that based on Z-Sequence is significant at the $p < 0.05$ level.

TR_{S100}, which suggests that when subtle differences in TR can be reflected, TR plays a crucial role in evaluating results. Therefore, we adopt this parameter in later experiments involving CR_{S100}. We compare r_s based on CR_{S100} to those based on TR_{S100} and CR_{S4} in Table 13.

Generally, all of the metrics based on CR_{S100} have better correlations with user satisfaction compared to those based on either TR_{S100} or CR_{S4}. For one thing, it confirms that TR and IQ are two different dimensions of an image and that both of them play an important role in the measurement of general relevance of image results. For another, most of the metrics based on CR_{S100} outperform the corresponding metrics based on CR_{S4}, which also certifies that fine-grain scales are more suitable for image search evaluation than traditional coarse-grain scales.

Finally, we attempt to combine all of the factors that we have discussed and calculate the Spearman's rho between two-dimensional metrics based on CR_{S100} and user satisfaction. Table 14 shows the results. The improvements are significant on all of the metrics, and the best row-based integration method for different metrics are almost consistent with results in Table 12. This result also verifies the benefits of considering row-based adaption, image quality, and fine-grain scales. As

Table 15. Spearman's Rank Correlation Coefficients (r_s) Between User Satisfaction and Offline Metrics Based on Different Annotations

	TR_{S4}	Row-ANT	Page-ANT	IQ	CR_{S4}	TR_{S100}
CG	0.277*	0.407*†	0.405*†	0.239*	0.379*†	**0.464***†
DCG@10r	0.269*	0.392*†	0.401*†	0.238*	0.365*†	**0.451***†
RBP (0.99)	0.330*	0.415*	0.401*	0.283*	0.419*†	**0.429***†
RBP (0.5)	0.211*	0.368*†	**0.401***†	0.156*	0.249*	0.302*†
ERR	0.193*	0.365*†	**0.401***†	0.140	0.221*	0.286*†
AVG	0.317*	0.419*†	0.405*	0.256*	0.411*†	**0.450***†

*Correlation is significant at the $p < 0.01$ level.
†Difference is significant at the $p < 0.05$ level, compared to the same metric based on TR_{S4}.

a result, metric CG with Row-MAX integration achieves the highest correlation ($r_s = 0.404$), followed by RBP ($p = 0.99$) and AVG with Row-MAX integration ($r_s = 0.403$), and it is the best result that we have achieved thus far.

5.5 Hard Queries

Considering the skewed distribution of topical relevance in our dataset, we further verify our findings on "hard" queries to eliminate accidental factors. We have not collected the user's feedback of each query's difficulty in the user study (Stage I), so we distinguish the query difficulty according to the relevance of image results returned by the search engine. In detail, the queries are sorted by the average topical relevance scores (S4) of image results in a descending order, and the last 25% are considered as hard queries. We recalculate the offline metrics and Spearman's rank correlation coefficients with user satisfaction based on different annotations. The results are shown in Table 15. Note that we do not include the metric "MAX" here to avoid mathematical exceptions, and metrics are calculated based on "Z-Sequence" when using TR_{S4}, IQ, CR_{S4}, and TR_{S100}.

The overall results are consistent with our findings regarding to RQ2 through RQ4. If we only consider the traditional four-level relevance, all of the metrics here have better correlations with user satisfaction than those of all queries, which indicates that offline metrics and TR have better discriminative power. This is not surprising, as we select the queries that have fewer relevant images. Row- and page-based TR outperforms the item-based one significantly in most cases. Moreover, the improvements in hard queries are greater than those in all queries in the aspect of absolute values. This emphasizes the impacts of the context in a row or a page in the user's perception of relevance. Comparing IQ and TR, we find that metrics using IQ alone fail to perform as well as those using TR alone, which seems contrary to the results in Table 9. We assume that users care more about TR when the image results are not so relevant—in other words, the query is hard—whereas users tend to pay more attention to IQ of each item if most of the images are highly relevant. Meanwhile, considering the metrics calculated based on combined relevance, we can still conclude that combining IQ and TR can improve the performance of offline metrics at a significant level. Furthermore, fine-grain relevance scale benefits offline metrics a lot because it further improves the discriminative power of TR and better reflects the user's perception of relevance.

Last but not least, we also find that metrics like CG and RBP (0.99) have better correlations with user satisfaction than metrics like ERR and RBP (0.5). This suggest that even in hard queries, users are still patient during examination and focus more on the gains.

Table 16. Comparison of Different Annotation Tasks

	Data Size	Type	Assessors per Unit	Platform	Instruction
Item TR_{S4}	79,337	Query-image	3	Baidu Zhongbao	Detailed instructions and examples
Item IQ	54,377	**Single image**	3	Baidu Zhongbao	Detailed instructions and examples
Row TR_{S4}	11,190	Query-image	3	Baidu Zhongbao	Explanations for each level
Page TR_{S4}	2,238	Query-image	3	Baidu Zhongbao	Explanations for each level
Item TR_{S100}	79,337	Query-image	**5**	Chinacrowds	Simple instructions

Table 17. Cost and Reliability of Different Crowdsourcing Annotation Tasks

	Total Cost	Time (Day)			Krippendorff's α
		Interface	Training	Main	
Item-based TR_{S4}	\$1,512	–	5	10	0.619
Item-based IQ	\$2,263	–	10	10	0.503
Row-based TR	\$491	10	5	10	0.788
Page-based TR	\$148	10	5	5	0.830
Item-based TR_{S100}	\$3,100	5	–	4	0.506

Table 18. Comparison of Costs and Performance Among Several Top Results

Metric	r_s	Annotation	Money	Time (Main/Total)
CG (Row-MAX)	0.404	Item-based TR_{S100} & Item-based IQ	\$5,363	14/29
CG (Row-MAX)	0.366	Item-based TR_{S100}	\$3,100	4/9
RBP (0.99)	0.345	Item-based TR_{S4} & Item-based IQ	\$3,775	20/35
RBP (0.99)	0.315	Item-based IQ	\$2,263	10/20
MAX	0.274	Row-based TR	\$491	10/25
RBP (0.99)	0.211	Item-based TR_{S4}	\$1,512	10/15

6 DISCUSSION

We mainly investigate how offline metrics calculated at different annotations align with user satisfaction in the last section. In this section, however, we first give a review of our annotation tasks through crowdsourcing platforms (Table 16) and the efficiency of these methods (Table 17). Then, we further discuss the cost and performances of several top results (Table 18).

6.1 Overview of Annotation Tasks

We exploited two popular crowdsourcing platforms in China: Baidu Zhongbao (in-house crowdsouring) and Chinacrowds (open crowdsourcing). For four-level annotation tasks, we employed Baidu Zhongbao because they had some experience in image annotation equipped with existing platforms and the work flow for annotation. In addition, we required that each annotation unit be annotated by three different assessors to further guarantee the annotation quality. However, we chose Chinacrowds, the more flexible open crowdsourcing platform rather than the in-house

crowdsourcing due to the specific design of both the interface and quality checks for the S100 annotation task. Meanwhile, we collected judgments from five different assessors to guarantee the reliability of annotation results. TR annotation was query dependent, so the query co-occurred with image results during annotation, whereas IQ was query independent, so the annotator could only make judgments according to the single image itself. The complexity of instructions also differed. We offered the most detailed explanations of each level and corresponding examples for the IQ annotation task. We also gave detailed instructions and examples when collecting S4 item-based TR annotations. As for row- and page-based TR, we gave the explanations of each level but without examples. And for the S100 TR annotation task, we only gave the simplest general instructions without any further explanations or examples, as we would like to gather more fine-grain perception of topical relevance from annotators.

6.2 Efficiency of Annotation Tasks

The in-house crowdsourcing company usually has the process of pre-training before the main task to make sure that their workers have fully understood the requirements, which is a bit time consuming, as Table 17 shows. To be detailed, in our tasks, the company selected some samples at random from the whole dataset and distributed them to the workers for annotating. Then they provided us with results of these samples. If we were not satisfied with the accuracy of these results, they would re-confirm our requirements and re-train their workers according to our feedback until the accuracy became acceptable. The time cost of training is highly related to the complexity of tasks and instructions. For example, the IQ annotation task was the most difficult because judgment of quality is a bit more subjective than that of TR. Although we provided the most detailed instructions along with examples, we still communicated with the manager several times to guarantee the accuracy during the training process, so the time for training was double. After the pre-training, the time to complete the main task is more related to data size. Thus, it took less time to complete the main task of page-based relevance annotation. Yet it barely took training time on the open crowdsourcing platform. Moreover, the open crowdsourcing seemed to complete tasks faster, considering the large number of flexible workers. However, we needed to design the interface and the quality checks by ourselves when utilizing the open crowdsourcing platform. As a result, it took us about 5 days to prepare and release the tasks but only 4 days to gather all annotation results. Note that we also needed to stitch image items and made some adjustments based on Baidu's existing framework for row- and page-based annotation tasks because there was no previous work on the annotation of row- and page-based relevance. In fact, it took about 10 days to communicate with the company, debu,g and put these adjustments into practice.

We have reported Fleiss's κ of four-level annotation tasks in the former section, which all reach moderate agreements [27]. However, Fleiss's κ can hardly reflect the reliability of S100 annotations due to the fine-grain scales. Therefore, we look at Krippendorff's α [20] instead.[16] As Table 17 shows, S100 annotations by open crowdsourcing has fair reliability compared to four-level item-based topical relevance and image annotations by in-house crowdsourcing.

The row- and page-based relevance annotation tasks cost less money because of the much smaller data size. It cost a bit more for IQ annotation compared to S4 TR due to the difficulty of the annotation task. The money cost of S100 annotation was the highest, but we collected scores from five different workers for each query-image pair.

[16]To be specific, we treat S100 as 101-level ordinal data.

6.3 Comparison of Costs and Performances

Furthermore, we select the metrics of top correlations with user satisfaction on the condition of different annotations and discuss their costs and performance (Table 18). We add up the time of different annotations, including the time of interface preparation, pre-training, and main task complement, and also list the time of completing main task separately since it is influenced by the data size. The highest correlation is given by two-dimensional metrics with Row-MAX integration based on a combination of S100 relevance and IQ annotations, but the added-up cost (money, time) is rather high at the same time. To take one step back, if we only consider S100 relevance annotations, the metrics performance would be sacrificed a little, but the cost, including both money and time, would be significantly reduced. In particular, we could achieve the second top correlation with the least time cost. It is worth mentioning that S100 relevance annotations outperform the combination of S4 TR and IQ annotations in correlation with user satisfaction yet with smaller cost both in money and time (the second and third row in Table 18). This suggests that S100 relevance annotating both S4 TR and IQ. We also find that IQ annotation can lead to better performance than S4 TR in our experiments, which might result from the imbalanced distribution of TR. But the cost of IQ annotation is higher considering the task difficulty and complexity. Considering S4 TR alone, row-based annotation is a more effective method than traditional item-based methods in that it can contribute to better correlation while costing much less money. However, some limitations on row-based relevance annotation exist. For example, it is difficult to combine quality with row-based relevance, and if the layout of images changes on SERPs, the annotated row-based relevance can hardly be re-used. These limitations are left for future work.

7 CONCLUSIONS AND FUTURE WORK

Search engine evaluation is essential in both academic and industrial IR research, and relevance annotation is a fundamental part of offline system effectiveness evaluation. Although image search engines show results in a different way than general web search, the impact of result annotations and the performance of offline metrics in the image search scenario are still under-investigated. To shed light on this research question, we design a two-stage data collection procedure and investigate how offline evaluation metrics align with user satisfaction on the condition of different data annotations. In our work, user satisfaction is considered as the golden standard for search performance evaluation. We collect explicit user satisfaction feedback through a lab-based user study in the first stage. Considering the factors that probably affect user perception, besides traditional four-level topical relevance of each image item, we gather large-scale image quality, row-based, page-based, and fine-grain TR annotations via both in-house and open crowdsourcing platforms. By looking at Spearman's rank correlation coefficients between user satisfaction and offline metrics calculated at different annotations, we compare the impact of different factors and attempt to figure out the efficient annotation methods and the suitable offline metrics for image search evaluation.

Centered on our research questions, we summarize our findings as follows. The highly imbalanced distribution of TR judgments in the traditional four-point scales weakens the discriminative power of TR and further limits the performance of offline metrics from the perspective of correlations with user satisfaction (RQ1). Our analysis confirms the important roles of the context of other results on the SERP (RQ2) and IQ (RQ3) for image search system evaluation. Contrary to TR, IQ is a query-independent factor and has the different distribution. Combining TR and IQ judgments can improve the performances of offline metrics significantly. In particular, the role of IQ changes with different search intents. Compared to item-based relevance, the row-based relevance is more suitable to reflect a user's perception when examining image results, whereas page-based

relevance is sometimes too broad and ambiguous. Fine-grain topical relevance has a smoother distribution and reflects more subtle differences in TR among images, and therefore it improves the performance of offline metrics significantly (RQ4). The highest Spearman's rank correlation coefficient with user satisfaction that we have achieved is 0.404 with a combination of fine-grain TR, IQ, and row-based integration.

For further discussion about the costs and performance of annotations, fine-grain (S100) TR annotation via the open crowdsourcing platform turns out to be a rather competitive choice, as it contributes to the second best performance and even outperforms S4 CR at a lower cost in both time and money. Compared to item-based relevance annotation, the row-based method seems to be more efficient with some better performance and has a much lower monetary cost.

Our study is the first to investigate the impact of different data annotations through crowdsourcing and correlations between offline metrics and user satisfaction in the image search scenario. The results provide insights for image search evaluation in aspects of both data annotation ways and offline metrics. Certainly, there are still some limitations to our work that we would like to list as our future work directions. User satisfaction is regarded as the golden standard for evaluation, but the gap between laboratory experiment environment and practical search scenarios would cause some bias when collecting a user's feedback. For example, practical image search users sometimes become impatient, whereas in our user study, participants tend to be more patient and examine numbers of images. In addition, participants in our experiment are all undergraduate students due to the resource constrains. More practical experimental designs that have a wide coverage, like field study [9, 21, 24, 51], will be investigated in the future. We only focus on query-level satisfaction evaluation in this article, and we would like to investigate session-level evaluation based on larger-scale and more practical data. Although we have observed the impact of the context factor and found that metrics based on row-based relevance annotations are more efficient than those based on item-based relevance, it is hard for us to combine the row-based relevance with other annotations. The reusability of row-based annotation is under-investigated. In this article, we mainly use row-based integration to consider the context in a row, which is a relatively naive simulation. In addition, we consider simple heuristic methods to combine TR and IQ, which might limit the performance of these factors. More effective combination methods are left for future work. In this article, we mainly investigate different annotation methods, which work as the basis of offline evaluation, while making only some slight adjustments to the existing metrics. The offline metric designed for the image search scenario is worth further study.

REFERENCES

[1] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 773–774.

[2] Omar Alonso and Stefano Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, Vol. 15. 16.

[3] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Information Processing and Management* 48, 6 (2012), 1053–1066.

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

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

[6] Ben Carterette. 2011. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 903–912.
[7] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge (CIKM'09)*. 621–630.
[8] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the International ACM SIGIR Conference*. 15–24.
[9] Karen Church, Mauro Cherubini, and Nuria Oliver. 2014. A large-scale study of daily information needs captured in situ. *ACM Transactions on Computer-Human Interaction* 21, 2 (2014), 10.
[10] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 terabyte track. In *Proceedings of the Text Retrieval Conference (TREC'04)*, Vol. 4. 74.
[11] C. W. Cleverdon and E. M. Keen. 1966. Aslib–Cranfield research project. *Factors Determining the Performance of Indexing Systems*, Vol. 1. College of Aeronautics.
[12] Jacob Cohen and Patricia Cohen. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.
[13] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M. Voorhees. 2015. *TREC 2014 Web Track Overview*. Technical Report. Michigan University, Ann Arbor, MI.
[14] Eli P. Cox III. 1980. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research* (1980), 407–422.
[15] Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, 162–170.
[16] Henry A. Feild, James Allan, and Rosie Jones. 2010. Predicting searcher frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 34–41.
[17] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, and Shipeng Li. 2011. The role of attractiveness in web image search. In *Proceedings of the 19th ACM International Conference on Multimedia (MM'11)*. 63–72.
[18] Abby Goodrum and Amanda Spink. 1999. Visual information seeking: A study of image queries on the World Wide Web. In *Proceedings of the ASIST Annual Meeting*, Vol. 36. 665–74.
[19] Qi Guo and Yang Song. 2016. Large-scale analysis of viewing behavior: Towards measuring satisfaction with mobile proactive systems. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16)*. 579–588.
[20] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1, 1 (2007), 77–89.
[21] Jiyin He and Emine Yilmaz. 2017. User behaviour and task characteristics: A field study of daily information behaviour. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, New York, NY, 67–76.
[22] Mehdi Hosseini, Ingemar J. Cox, Nataša Milić-Frayling, Gabriella Kazai, and Vishwa Vinay. 2012. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Proceedings of the European Conference on Information Retrieval*. 182–194.
[23] Kalervo Järvelin and Jaana Kekäläinen. 2002. *Cumulated Gain-Based Evaluation of IR Techniques*. Vol. 20. ACM, New York, NY.
[24] Diane Kelly and Nicholas J. Belkin. 2004. Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 377–384.
[25] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement? In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'18)*. 805–814.
[26] Dmitry Lagun, Chih Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.
[27] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.
[28] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 493–502.
[29] Cheng Luo, Yiqun Liu, Tetsuya Sakai, Fan Zhang, Min Zhang, and Shaoping Ma. 2017. Evaluating mobile search with height-biased gain. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 435–444.

[30] Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing*.

[31] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems* 35, 3 (2017), 19.

[32] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, et al. 2016. When does relevance mean usefulness and user satisfaction in web search? In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. 463–472.

[33] Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2011. Crowdsourcing blog track top news judgments at TREC. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. 23–26.

[34] Alistair Moffat, Paul Thomas, and Falk Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*. 659–668.

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

[36] Howard R. Moskowitz. 1977. Magnitude estimation: Notes on what, how, when, and why to use it. *Journal of Food Quality* 1, 3 (1977), 195–227.

[37] Neil O'Hare, Paloma De Juan, Rossano Schifanella, Yunlong He, Dawei Yin, and Yi Chang. 2016. Leveraging user interaction signals for web image search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. 559–568.

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

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

[40] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On fine-grained relevance scales. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 675–684.

[41] Tetsuya Sakai. 2018. Conducting laboratory experiments properly with statistical tools: An easy hands-on tutorial. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 1369–1370.

[42] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 473–482.

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

[44] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 2126–2144.

[45] David J. Sheskin. 2003. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press, Boca Raton, FL.

[46] Mark D. Smucker and Charles L. A. Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 95–104.

[47] Mark D. Smucker, Gabriella Kazai, and Matthew Lease. 2012. *Overview of the TREC 2012 Crowdsourcing Track*. Technical Report. Texas University at Austin School of Information.

[48] 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 *Proceedings of the 22nd International Conference on World Wide Web*. ACM, New York, NY, 1201–1212.

[49] Eero Sormunen. 2002. Liberal relevance criteria of TREC: Counting on negligible documents? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 324–330.

[50] Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Alexandru Lucian Ginsca, Adrian Popescu, Yiannis Kompatsiaris, and Ioannis Vlahavas. 2015. Improving diversity in image search via supervised relevance scoring. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, New York, NY, 323–330.

[51] Xu Sun and Andrew May. 2013. A comparison of field-based and lab-based experiments to evaluate user experience of personalised mobile devices. *Advances in Human-Computer Interaction* 2013 (2013), 2.

[52] Rong Tang, William M. Shaw, and Jack L. Vevea. 2010. Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science and Technology* 50, 3 (2010), 254–264.

[53] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 565–574.

[54] Reinier H. van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. 2009. Visual diversification of image search results. In *Proceedings of the 18th International Conference on World Wide Web*. ACM, New York, NY, 341–350.

[55] Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval*. Vol. 1. MIT Press, Cambridge, MA.

[56] 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 *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM'18)*. 655–663.

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

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

[59] 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 *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 615–624.

[60] Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M. Jose, and Leif Azzopardi. 2013. Crowdsourcing interactions: Using crowdsourcing for evaluating interactive information retrieval systems. *Information Retrieval* 16, 2 (2013), 267–305.

Received August 2018; revised November 2018; accepted January 2019