The Influence of Image Search Intents on User Behavior and Satisfaction

Zhijing Wu[†], Yiqun Liu[†*], Qianfan Zhang[‡], Kailu Wu[‡], Min Zhang[†], Shaoping Ma[†]
†Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
‡Institute for Interdisciplinary Information Sciences,
Tsinghua University, Beijing 100084, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

Understanding search intents behind queries is of vital importance for improving search performance or designing better evaluation metrics. Although there exist many efforts in Web search user intent taxonomies and investigating how users' interaction behaviors vary with the intent types, only a few of them have been made specifically for the image search scenario. Different from previous works which investigate image search user behavior and task characteristics based on either lab studies or large scale log analysis, we conducted a field study which lasts one month and involves 2,040 search queries from 555 search tasks. By this means, we collected relatively large amount of practical search behavior data with extensive first-tier annotation from users. With this data set, we investigate how various image search intents affect users' search behavior, and try to adopt different signals to predict search satisfaction under the certain intent. Meanwhile, external assessors were also employed to categorize each search task using four orthogonal intent taxonomies. Based on the hypothesis that behavior is dependent of task type, we analyze user search behavior on the field study data, examining characteristics of the session, click and mouse patterns. We also link the search satisfaction prediction to image search intent, which shows that different types of signals play different roles in satisfaction prediction as intent varies. Our findings indicate the importance of considering search intent in user behavior analysis and satisfaction prediction in image search.

KEYWORDS

Search intent; field study; user behavior; user satisfaction

ACM Reference Format:
Zhijing Wu[†], Yiqun Liu[†], Qianfan Zhang[‡], Kailu Wu[‡], Min Zhang[†], Shaoping Ma. 2019. The Influence of Image Search Intents on User Behavior and Satisfaction. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291013

*Corresponding author

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3291013

Table 1: An example of diverse search intents under the query "Ready Player One" (an American film) on image search platform.

Query	Ready Player One	
Intent	**Participant 1:** I have made an appointment with my friends to go to the movie theater this weekend. Before that, I want to browse some related pictures about the popular film "Ready Player One".	**Participant 2:** I am preparing slides for a presentation about American film. I want to download an image of the poster of the film "Ready Player One" that I have seen in the movie theater.
why	The user is going to freely browsing some pictures in leisure time (*Entertain*). The search is driven by daily-life needs (*Daily-life*).	The user is going to find and download pictures for further use (*Locate*). The search is driven by work requirements (*Work&Study*).
what	The search goal is broad, including posters, stills, actors and so on (*General*). Before search, the user did not know how the image content looks like (*Navigation*).	The search goal is specific, only searching for the poster (*Specific*). Before search, the user knew how the image content looks like (*Mental Image*).

1 INTRODUCTION

User intent understanding has become a hot topic in the field of Web search since it helps provide better search experiences for users. One of the most popular search intent taxonomies was introduced by Broder [3], which groups Web search users' intents into three categories: *informational*, *transactional* and *navigational*. Kofler et al. [19] showed that a user information need is composed of two dimensions: "what" dimension (what users are searching for) and "why" dimension (why users search). It demonstrates that users have diverse information needs during the search processes. However, only a few of search intent studies have been made specifically for the image search scenario, in which the content that users search for, the presentation of search results, and user behavior patterns are rather different from general Web search [30, 41]. Similar with general Web search, even if users submit the same query into an image search engine, the underlying search intents can be rather different. Table 1 shows an example of different users' search intent descriptions under the same query "Ready Player One" (an American film) sampled from the image search behavior log in our field study (will be described in Section 3). User 1 wanted to get information by browsing various images about the film, while the user 2 expected a concentrated search object and needed to download images for further use. The search goal and purpose vary

in these two scenarios. Consequently, users may have different search behavior and be satisfied with different image results. It is important to investigate the image search intents behind queries for a better understanding of user behavior and satisfaction.

In recent studies, user behavior data has been widely used for task classification and satisfaction prediction. Many search intents are triggered by the landing pages that users browse right before their search actions [5]. Therefore, the concept of predicting users' search intents based on their browsing behaviors is proposed. Several studies focus on the relationship between search task and interaction with image search engines. Park et al. [29] categorized queries using two orthogonal taxonomies (subject-based and fact-based) and found that there are a number of differences in search behavior across query types. For example, some task types are associated with exploratory, browsing-style behavior, while with other intent types users exhibit a more focused search. These search behavior can also provide a better understanding of user satisfaction with hand-crafted aggregated features [9] or modeling user behavior with original action sequences [11, 12]. With respect to "why" based intent taxonomies in image search, Xie et al. [40] propose a user intent recognition method based on user interactions at the early stage of search sessions, which demonstrates that users interact with image search engines in different ways as intent varies.

Since the search scenario and interaction interface are quite different between general Web search and image search [1, 30, 41], existing work is not enough to thoroughly understand users' search intent in image search. Previous studies show that search intent can affect users' query reformulation behavior and interaction with search results. At the meantime, such user behavior is a strong implicit feedback for search satisfaction. To our best knowledge, there exits almost none work in image search trying to investigate the relationship among intent, behavior, and satisfaction, which may provide useful opinions for satisfaction prediction and search result evaluation in various search scenarios. Meanwhile, most existing studies focus on a particular type of search intent taxonomy. Comparisons among different taxonomies, especially those between "why" based and "what" based taxonomies, are not involved. This motivates our following research questions:

- **RQ1:** How does image search behavior change with user intent in different intent taxonomies?
- **RQ2:** What factors affect users' perception of satisfaction across different image search intent taxonomies?
- **RQ3:** How do different types of signals (e.g. click, mouse movement, users' explicit feedback) perform in the prediction of user search satisfaction across different search intents in different search intent taxonomies?

Most of user search intent related studies are based on practical search log analysis [3, 29, 33, 34] or lab study [24, 40]. Log data from the commercial search engine can provide large-scale and practical search behavior data for researchers, while it is difficult to get explicit feedback (e.g. search satisfaction) from users directly. In lab studies, researchers collect users' feedback right after each designed search task. However, these studies are performed under a controlled environment and may not reflect users' true search intent or practical behavior patterns. In this paper, we conduct a field study for one month, during which we log participants' daily image search activities, and they are asked to self-annotate their satisfaction and search behavior motivations (e.g. intent description, evaluation criterion, the reason for click). This field study can provide us with more accurate annotations and more practical search behavior data.

Based on image search data collected from the field study, we also employ external experts to categorize search sessions using four orthogonal intent taxonomies (i.e. *Locate & Learn & Entertain* [40]; *Work & Daily-life* [15]; *Specific & General* [22]; *Mental Image & Navigation* [24]. See Section4 for details). We try to answer the above three research questions and gain a better understanding of users' image search behavior with this dataset.

To summarize, the main contributions of this work are as follows:

- We construct a field-study based image search behavior data set to investigate users' daily image search intents. The data set contains practical behavior data, data labels from user themselves, and image relevance annotations collected through crowdsourcing [1]. It can provide researchers with a more reliable and realistic view than traditional lab study and large-scale log analysis on how search intents are associated with user behavior and satisfaction.
- We investigate the differences in user behavior under different search intent using existing taxonomies according to both what they are searching for and why they search. We also focus on the relationship among intent, behavior, and satisfaction. It indicates the importance of considering search intent to better understand user behavior in image search and the necessity of designing evaluation metrics respectively for different search tasks.

The rest of this paper is organized as follows. We review the related work on intent taxonomy, user behavior and satisfaction in Section 2. In Section 3 and Section 4, we introduce the details of our field study and search intent taxonomies. Users' implicit and explicit feedbacks are investigated in Section 5 to answer **RQ1**&**RQ2**. Furthermore, we try to predict search satisfaction using different features under certain search intent in Section 6 to answer **RQ3**. Finally, we discuss the conclusions and some limitations.

2 RELATED WORK

Search intent taxonomy: Search intent behind queries can be divided into two dimensions: "what" users are searching for and "why" they search [2]. One of the most popular search intent taxonomies was proposed by Broder [3] for text-based Web search: *informational*, *navigational*, and *transactional*. Based on this taxonomy, several intent taxonomies were proposed for general Web, image, and video search [6, 10, 20, 24, 40]. Since search task is defined as activities that people attempt to accomplish in order to keep their work or life moving on [22], search tasks are widely studied in work task level and search task level [4, 36, 37, 39]. In image search, one of the most important intent taxonomies is proposed by Lux et al. [24], which categories search intent into navigation, transaction, knowledge orientation and mental image. Recently, Xie et al. [40] focused on why people search and showed that intents can be grouped into three classes: Explore/Learn, Entertain, and Locate/Acquire in image search. Mitsui et al. [27] showed that information seeking intentions can be predicted with a simple classification model.

Only a few of search intent related studies have been made specifically for the image Web search scenario. In this paper, we try to compare different types of "why" based and "what" based taxonomies in image search.

[1]The data is now available at http://www.thuir.cn/group/~YQLiu/

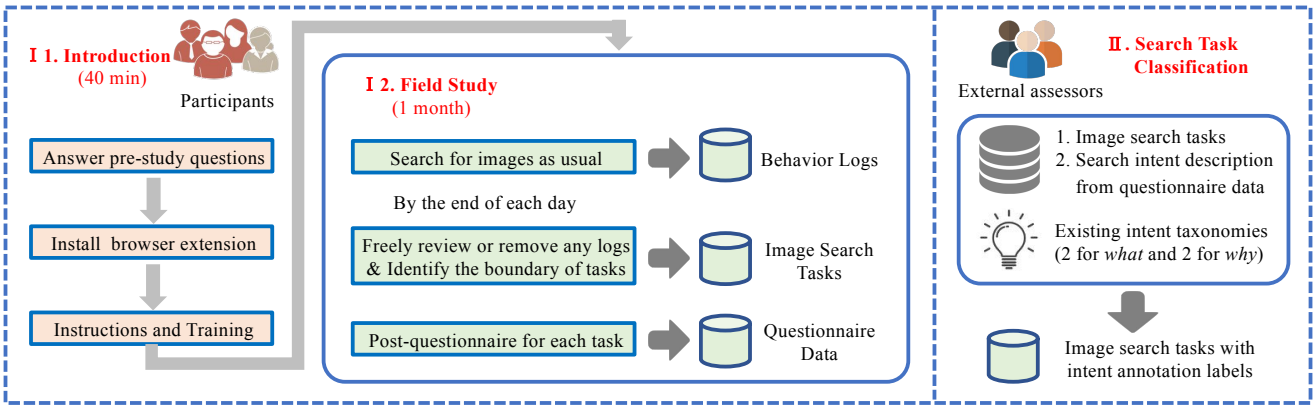

Figure 1: We can get practical image search behavior data as well as search feedback (e.g. search intent description, satisfaction) from user themselves through the field study (part I). External assessors are employed to classify the search tasks according to existing search intent taxonomies through the task annotation procedure (part II).

User behavior in image search: Previous studies have conducted large-scale log analysis to investigate the user behavior in image search and compare the differences with general Web search [1, 30, 38]. They found that image search users usually submit shorter query strings and their selections of query terms are more diverse. Many features such as session length, browsing depth, and query reformulation patterns are also measured to characterize the general behavior of image search users [16, 28]. Eye-tracking devices are also used to get more detailed observations about how users interact with the search result. Xie et al. [41] found a middle-position bias in users' image search examination patterns through a laboratory eye-tracking study. Underwood and Foulsham [35] investigated how visual saliency affects users' eye movements. Park et al. [29] linked the user behavior to search intent. They analyzed a large-scale query log from Yahoo Image Search to investigate user behavior toward different query types and identified important behavioral differences across them. Rha et al. [31] found that there are some differences in query reformulation types following different search intentions through an observational study.

These studies often investigate user behavior based on lab studies or large-scale log analysis. There are few researches that study users' image search behavior based on field study, through which we can collect practical behavior data as well as data labels from user themselves. It can provide us with a more reliable and realistic view than traditional lab study and large-scale log analysis on user behavior understanding.

User satisfaction: User satisfaction is defined as the fulfillment of information requirement to measure users' subjective feelings about search process [17]. User behaviors have been extensively used to predict user satisfaction. Hassan et al. [13] demonstrated that a query-based model (using the relationship between users' current and next query) can indicate satisfaction more accurately than click-based models. Kim et al. [18] utilized three measures of dwell time for predicting click-level satisfaction. Guo et al. [9] showed that fine-grained interactions (e.g. mouse movements and scrolling) can provide additional clues for better predicting task-level satisfaction.

However, to the best of our knowledge, few existing studies build models for search tasks with different intents respectively when predicting user satisfaction in image search scenario. We try to link the user satisfaction prediction to image search intent in this paper.

3 FIELD STUDY

In this section, we describe the details of our field study (which was designed based on He's field study on Web search [14]) and the dataset we used throughout this paper. The procedure of the filed study is shown in Figure 1(I).

Introduction (40 minutes, Figure 1(I1)). We invited participants to the lab with their own laptops. They were asked to fill a pre-questionnaire to collect demographic information and the usage of image search engines. Meanwhile, we installed the web browser plug-in on their laptops to record their daily image searching activities. After an introduction of the study procedure, we began to train the participants to get familiar with the platform. They were instructed to complete one recently engaged search task with the web browser plug-in and annotate the task on the annotation platform. After ensuring that the participants had been familiar with the study procedure, they were told to go back home and use their laptops as usual.

Recording and annotating (1 month, Figure 1(I2)). The field study lasted for one month, during which participants' image search activities were recorded automatically by the web browser plug-in (see Section 3.1 for details). By the end of each day, they were required to log into the annotation platform to examine their image search logs. They could remove any logs that they did not want to share. After that, they needed to identify the queries that were submitted for the same search task and fill a post-questionnaire for each task as search feedback (see Section 3.2 for details). Meanwhile, we manually examined the post-questionnaires submitted by participants. Reminder email would be sent to a participant if we found that there were problems with his/her feedback data (e.g. He/She did not complete the task questionnaire in time).

Ending the study. After one month, participants were informed to uninstall the web browser plug-in on their laptops and they were paid according to the number of post-questionnaires they submitted. Finally, they were required to answer several questions about the experience during the field study (e.g. What do you think of the annotation interface? How many search tasks have you deleted from the search logs?). We find that participants felt satisfied with our study design and they rarely removed any search logs. Through this study procedure, we can get image search log data reflecting participants' true daily information needs. Asking participants to

Table 2: Post-questionnaire questions and descriptions.

	Attribute	Question	Value and Description
Task	Intent (TI)	What's the objective of this image seeking activities?	open question
	Satisfaction (TSAT)	How satisfied are you with the whole image seeking experience and search results for this task?	① unsatisfied → ⑤ very satisfied
	Evaluation Criterion (EC)	During this search process, which characteristic of image results would affect your search satisfaction? Please select at least two most important characteristics.	*A*.relevance; *B*.content diversity; *C*.visual diversity; *D*.aspect ratio; *E*.size; *F*.watermark; *G*.aesthetics; *H*.original website; *I*.other____
Query	Satisfaction (QSAT)	For each query submitted in this task, How satisfied are you with the image seeking experience and search results?	① unsatisfied → ⑤ very satisfied
Clicked Image	Relevance (CIR)	For each image result clicked in this task, how do you rate its relevance to this query?	① irrelevant → ⑤ highly relevant
	Usefulness (CIU)	For each image result clicked in this task, how do you rate its usefulness for your search task objective?	① useless → ⑤ highly useful
	Reason for Click (CR)	Why do you click on this image result? Please select at least two most important reasons.	*A*.relevance; *B*.surrounding text; different with surrounding images in *C*.content; *D*.visual presentation; *E*.rank position; *F*.utility; *G*.aesthetics; *H*.just interested in it; *I*.other____

divide search tasks by themselves is more accurate than traditionally separating tasks when the time between consecutive actions exceeds 30 minutes [29]. Meanwhile, we collect searching feedback from participants themselves directly.

3.1 Search Logging

We develop a plug-in for the Chrome browser to automatically log participants' daily activities with image search engines. Since the plug-in is only installed on Chrome browser, participants are told to use Chrome for searching images during this field study. The information we recorded are as follows:

- **Keyboard activities.** We record the query that participants input and submit to the search engine.
- **HTML.** We save the URLs and HTML contents that participants have browsed, which include search engine result pages (SERPs) and image result preview pages (the page that is shown after clicking on an image result, which contains an enlarged preview of the image result).
- **Mouse activities.** The mouse activities include movement, scrolling, hover, and click events.
- **Timestamp.** All of the data above is associated with a timestamp, with which we can calculate the dwell time on SERPs, first click time, mouse moving speed and so on.

3.2 Search Feedback

During the field study, participants needed to log in to our annotation website to complete the search feedback for their own search logs. The website supports the following operations: (1) review and remove logs; (2) identify search tasks; (3) fill post-questionnaire.

Review and remove logs. Participants can review the queries they have submitted to search engine and URLs of SERPs they have browsed at any time through our annotation website. Meanwhile, they are told to freely remove the queries that they do not want to share with us. Once a query is removed, all records related to this query will be removed from the logs. Allowing participants to remove the logs can make them feel free when searching on the Internet.

Identify search tasks. A search task is a trigger for users to consult a search engine with a set of (textual) queries [19]. Since we focus on the search intent behind a task rather than a query, participants were required to identify the queries belonging to the same task. To ensure that participants are clear about the concept of task, we asked them to think of several queries they had been recently submitted and practise task identifying during the pre-experiment training.

Post-questionnaire. Participants need to fill a post-questionnaire for each task after they identify the search tasks. We list the following information to help participants to complete the feedback: (1) queries and SERPs for each task; (2) clicked images for each query. The details of post-questionnaire are shown in Table 2. Task intent is the "immediate reason, purpose, or goal" that motivates a user to consult the search engine [19]. We encourage participants to describe the details of search intent as specific as possible. We provide several most possible options for evaluation criteria (EC) and the reasons for click (CR). If a participant select "content diversity" as evaluation criteria for a task, it means that he/she wants a more diversified result list in content during this search task. "Aspect ratio" indicates the ratio of width to height of image result. "Watermark" is a special mark contained in images that are used to stop people from copying them. Here we define the "utility" of image results as a synthesis of aspect ratio, size, and watermark. Participants are allowed to input other answers that are not involved in these options. In fact, we find that only a few participants input new answers in the collected data. They are "lazy" when filling the feedback questionnaire. Therefore, we mainly focus on the options we provide in the subsequent analysis. We use a 5-level satisfaction scale [23] (1: unsatisfied, 2: slight satisfied, 3: fair satisfied, 4: substantial satisfied, 5: very satisfied), a 4-level relevance scale [42] (1: irrelevant, 2: somewhat relevant, 3: fairly relevant, 4: highly relevant), and a 4-level usefulness scale [25] (1: not useful at all, 2: somewhat useful, 3: fairly useful, 4: very useful).

3.3 Participants and collected data

We recruited 50 participants, 23 females and 27 males, to take part in the field study. Their ages range from 18 to 36. 36 of them are

Table 3: The agreement (Fleiss kappa and the ratio of tasks that three assessors' annotations are the same) of search intent annotation and the distribution of different types of tasks according to majority vote.

	why dimension					*what* dimension			
	Locate/Learn/Entertain			Work&Study/Daily-life		Specific/General		Mental Image/Navigation	
Fleiss Kappa	0.608			0.718		0.484		0.586	
Consistent Ratio	0.607			0.802		0.622		0.717	
	Locate	**Learn**	**Entertain**	**Work&Study**	**Daily-life**	**Specific**	**General**	**Mental Image**	**Navigation**
Majority Vote	179	192	184	201	354	321	234	182	373

undergraduate or graduate students majoring in engineering and arts. 14 of them are postdoctoral researchers or staffs in the university. All the participants report that they use image Web search engines for studying, working or other daily purposes, so they are familiar with the basic usage of Web search engines.

Through the field study, we collected an image search dataset that contains 592 search tasks. Because of some problems (Only part of behaviors during the search task were recorded because the participant accidentally closed the web browser plug-in) in recording behavior logs, we filtered out some search tasks. Finally, we have 555 tasks, 2040 queries, 270,315 image results and 2,700 clicks. Participants submitted 3.68 queries per task and browsed 132 images (The number of images that were loaded on the SERP.) per query on average. Participants may click on an image result for an enlarged version or download the image. 1% images are clicked and the average number of clicks per query is 1.32.

4 SEARCH INTENT ANNOTATION

After collecting participants' search task behavior and explicit feedback in image search, we employed external assessors to make judgments for the search task intent of field study data according to existing intent taxonomies. Since it is difficult to ensure all participants have the same classification criteria, we do not ask participants themselves to annotate the task category. The results shows that external assessors can reach a moderate or higher agreement on intent classification with participants' search intent description.

4.1 Search Intent Taxonomy

Following previous work [15, 22, 24, 40], we adopt four search intent taxonomies in this study, two of which focus on *why* dimension (why users search) and the other two focus on *what* dimension (what users are searching for). Each of the four taxonomies is applicable to every search task. The example of each task type can be found in Table 1. Participant 1 conducted an *Entertain*, *Daily-life*, *General*, *Navigation* search task, while participant 2 conducted a *Locate*, *Work&Study*, *Specific*, *Mental Image* task. Firstly, we introduce the taxonomies of *why* dimension:

Locate, Learn, Entertain. Xie et al. [40] classify the search intents according to (1) whether the user's search behavior is driven by a clear objective, (2) whether the user needs to download the image for further use after the search process. Under *Locate* intent, users want to find images for further use. They already have some requirements for these images. Under *Learn* intent, users want to learn, confirm or compare information by browsing images. For example, the intent of one search task in our dataset is described as follows: I want to know what is "Passiflora edulis" and how it looks like. Under *Entertain* intent, users want to relax and kill time by freely browsing images. An example of *Locate*/*Entertain* is the search intent of participant 2/1 shown in Table 1.

Work&Study, Daily-life. Ingwersen and Järvelin [15] classify the search intents according to whether the search behavior is driven by work requirement or not. Since more than half of the participants in our field study are students, we adjust the "work requirement" to "work or study requirement". A *Work&Study* task refers to an activity people perform in order to fulfill their needs for jobs or courses.

Secondly, we describe the taxonomies of *what* dimension:

General, Specific. Li and Belkin [22] classify the search intents according to the goal of search behavior. For the *General* task, users want to get general information about a subject, while for the *Specific* task, users have explicit or concrete goals.

Mental Image, Navigation. Lux et al. [24] classify the search intents into *Mental Image* and *Navigation*. Under the *Mental Image* intent, the user knows how the image content looks like before search and looks for images to match to the mental image by visual analysis. Under *Navigation* task, the user knows the existence of the image, but its content is unknown.

4.2 Search Intent Annotation

We recruited twelve external assessors who were divided into four groups (three assessors in each group) to annotate the task intent type as Figure 1(II) shows. The assessors were from a commercial search engine company and were familiar with the intent classification task. Each group of assessors made annotations for 555 tasks according to one taxonomy introduced above. Before annotation, we gathered the assessors who were in the same group to introduce the taxonomy. After the introduction, they had 10 minutes to discuss the criteria of classification to ensure that they were clear about the criteria. Then they were shown the participant's intent description and the query list for each task and were asked to make judgments alone based on this information. It took about four hours to complete the annotation.

The agreement is shown in Table 3 to assess the reliability of annotation. We report the value of Fleiss' Kappa [8], which ranges from 0 to 1 (0-0.2: slight agreement, 0.2-0.4: fair agreement, 0.4-0.6: moderate agreement, 0.6-0.8: substantial agreement, 0.8-1.0: almost perfect agreement [21]). We also report the ratio of tasks where the three assessors' annotations are the same. We find that the classification results of four taxonomies can reach a moderate or higher agreement. It suggests that the search task can be classified by external assessors with users' intent description. Annotations according to taxonomies of *why* dimension, by contrast, are easier to reach high agreement than annotations according to those of *what* dimension. From the feedback of assessors, they feel rather difficult to tell *Specific* tasks and *General* tasks.

We use the majority vote of three assessors as the category label of search tasks. Finally, we conduct an image search dataset consisting of 555 tasks and each task has four category labels (corresponding to four intent taxonomies). We further collect the relevance

Table 4: Differences in user behavior with different search intents. Results in boldface are significant higher than that of other task types under the same taxonomy. "/*" indicates that statistical significance at $p\text{-}value < 0.01/0.05$ level (One-way ANOVA) among different task types of one taxonomy.**

	Behavior Feature	*why* dimension						*what* dimension					
		Locate	Learn	Entertain		Work &Study	Daily Life		Specific	General		Mental Image	Navigation	
Task	number of queries	**4.25**	3.45	3.35	*	4.09	3.44	-	2.91	**4.73**	**	3.38	3.79	-
	number of query terms	**11.0**	8.23	8.05	*	10.3	8.39	-	7.05	**11.9**	**	8.77	9.20	-
	number of unique query terms	**5.78**	4.85	4.73	*	5.58	4.85	-	4.02	**6.61**	**	4.90	5.20	-
	unique query terms ratio	0.68	0.72	**0.77**	**	0.71	0.73	-	0.73	0.72	-	0.72	0.72	-
Query Text	number of terms	**2.93**	2.68	2.57	**	**2.83**	2.68	*	2.76	2.73	-	**3.01**	2.65	**
	number of Chinese characters	**6.42**	6.17	5.75	**	**6.52**	5.88	**	**6.35**	5.96	*	**6.62**	5.97	**
Click	number of clicks	1.43	**1.45**	1.06	**	1.14	**1.45**	**	1.40	1.26	-	1.42	1.29	-
	first click time (s)	**13.0**	11.4	9.45	**	**12.5**	10.6	**	11.1	11.2	-	11.2	11.1	-
	last click to end time (s)	**31.9**	24.1	21.9	**	**32.5**	22.7	**	27.4	24.4	-	28.0	25.0	-
	min of click depth (row)	**5.98**	4.02	2.78	**	**5.14**	3.73	**	4.13	4.22	-	4.47	4.07	-
	max of click depth (row)	**13.7**	8.05	4.16	**	**10.4**	7.49	**	**9.48**	7.66	*	**9.91**	7.93	*
Dwell time	dwell time on SERP (s)	**32.1**	28.2	19.2	**	**29.5**	25.2	**	28.2	25.9	-	27.7	26.7	-
	dwell time on preview page (s)	**23.1**	18.1	12.1	**	**21.2**	15.8	**	**20.1**	15.8	*	19.7	16.9	-
Mouse	moving time ratio	**0.93**	0.88	0.84	**	**0.90**	0.87	**	**0.92**	0.86	**	**0.90**	0.88	**
	average of moving distance (pix)	**73.3**	73.2	58.3	**	**72.7**	66.0	**	67.7	68.4	-	70.6	67.3	-
	median of moving speed (pix/s)	301	**334**	327	*	302	**331**	**	297	**340**	**	325	321	-
	scrolling time ratio	**0.43**	0.37	0.31	**	**0.40**	0.36	**	**0.41**	0.35	**	0.37	0.37	-
	average of scrolling distance (pix)	**89.4**	86.2	75.6	**	**90.9**	81.3	**	79.7	**89.4**	**	83.5	85.3	-
	median of scrolling speed (pix/s)	**614**	594	544	**	**620**	571	**	564	**615**	**	564	**599**	*

scores of all image results using the methods introduced by Roitero et al. [32] in this dataset, which are not used in this paper.

5 IMAGE SEARCH WITH DIFFERENT INTENTS

In this section, we first compare the differences in user behavior (implicit signals) among search intents to answer **RQ1** (How does image search behavior change with user intent in different intent taxonomies). Then we analyze participants' explicit feedbacks of evaluation criteria to answer **RQ2** (What factors affect users' perception of satisfaction across different image search intent taxonomies). Since click is an important signal for users' satisfaction perception [7], we also analyze participants' explicit feedbacks of reasons for click to better understand users' search process.

5.1 Implicit Signals

We conduct one-way ANOVA on the 555 tasks to analyze the effect of search intents on each measure related to participants' search behavior. Table 4 shows the means of search behavior features and significance levels across different task types under the same taxonomy. We categorize the search behavior measures into 5 groups, which are related to the task, query text, click behavior, dwell time, and mouse behavior respectively. The latter four groups are query-level features, for which we report the average of all queries belonging to one certain type of tasks.

Task. We segment the query text (Chinese) and remove the stop words. The "query term" refers to how many words there are in the query text. Under the "*Locate/Learn/Entertain*" taxonomy, the search intent has a significant effect on task-related measures. The task length (how many queries are there in one task) is longer when users want to find images for further use (*Locate* type) or the search goal is general. Results are the same for the number of query terms and unique terms in one task. In *General* tasks, users have a broad search scope, which leads to more query terms in the whole task. Note that the unique query terms ratio is in a low level in *Locate* tasks, which indicates that users submit more duplicate terms when looking for images to download.

Query text. The query-level number of terms and Chinese characters in *Locate*, *Work&Study*, *Mental Image* queries are significantly larger than that in other query types. It indicates that these three types of tasks are more complex and challenging. Especially in *Mental Image* tasks, the mean of terms reaches 3.01. Users need to formulate longer and more complex queries to describe their mental images, which also puts forward higher requirements for search engines.

Click behavior. The search intent has a significant effect on click-related features when classifying tasks according to taxonomies of *why* dimension, while the differences do not reach statistical significance according to taxonomies of *what* dimension except the max of click depth (Click depth is defined as the row number of the clicked image because image search engines organize results by row). When users want to learn, confirm, or compare information from image results (*Learn* type) or download images (*Locate* type), they need to click on the results to examine the details of images on an enlarged version, which leads to a larger number of clicks. Since users can observe the thumbnails of image results on SERP, they will spend more time to decide which image to download and click, which leads to a longer time before the first click and a deeper depth of clicks. The *Work&Study* tasks show a similar tendency with *Locate* tasks. When users have a goal that needs to be fulfilled as responsible to jobs or courses, they are more "patient" with the search results and click on results ranked beyond the first 10 rows. Similar results are observed when users have specific search goals or mental image. They interact with results ranked deeper than users who are performing general or navigation tasks.

Dwell time. In image search, a preview page is loaded after users click on an image result, through which users can download full-size images, browse other results without going back to the SERP. We analyze the dwell time on SERP as well as preview page. These two signals indicate the explore behavior on SERP and preview page. The search intent has a significant effect on dwell time when classifying tasks by taxonomies of *why* dimension, while the differences do not reach statistical significance by taxonomies of taxonomies of *what* dimension. It indicates that users' explore behavior is dependent on why they search rather than what they are searching for.

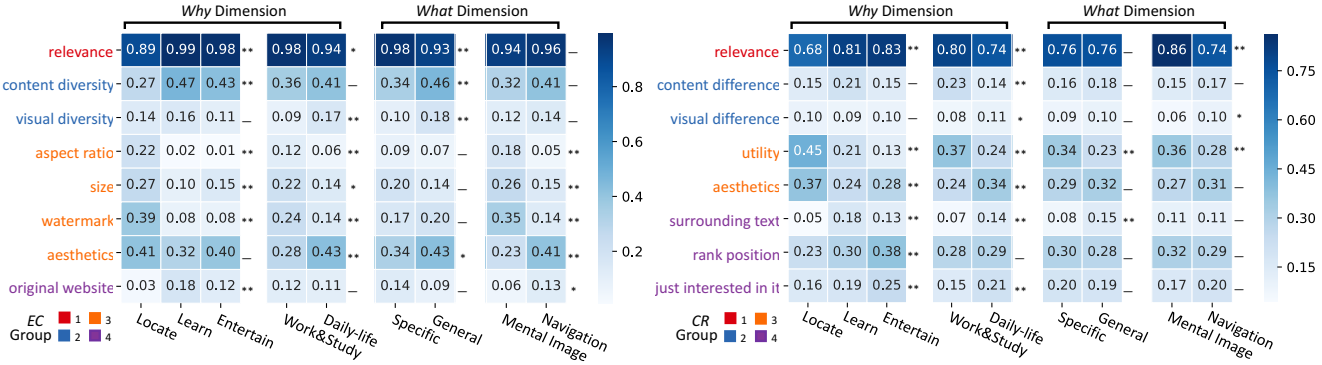

Figure 2: The distribution of Evaluation Criteria (EC, left) and Reasons for Click (CR, right) participants selected after search across different search intents. For example, the "0.89" in the top left corner means that participants select relevance as one of their *EC* in 89% of *Locate* tasks. The options of *EC* involve 4 groups: 1) relevance, 2) diversity, 3) quality, 4) website. The options of *CR* involve 4 groups: 1) relevance, 2) diversity, 3) quality, 4) other. "/*" indicates that statistical significance at $p\text{-}value < 0.01/0.05$ level (One-way ANOVA) among different task types of one taxonomy.**

Mouse behavior. We analyze the distance and speed of mouse movement and scrolling. We only list part of significant results in Table 4. The ratio of moving/scrolling time can reach 0.93/0.43 at most. In *Learn*, *Work&Study*, and *Specific* tasks, the mouse activity ratios are significantly higher than other tasks under the same taxonomy, which may provide insights for search intent classification.

5.2 Explicit Signals

We use the post-questionnaire data of our field study to analyze what factors affect user satisfaction and click behavior. Participants were allowed to select at least two criteria from the options we provided (listed in Table 2) or input factors we did not provide, while very few participants inputted new factors. Therefore, we only focus on the eight factors that we provide as options. We divide the evaluation criteria into 4 groups: relevance, diversity, quality, and website. Since users cannot get the information about the images' original website before they click on the image, we did not involve website as the reason for click. The four groups of reasons for click are relevance, diversity, quality, and other. We report the distribution of evaluation criteria, reasons for click and statistical significance level of One-way ANOVA as Figure 2 shows.

We find that relevance is the major concern when perceiving satisfaction, followed by the quality, diversity, and original website. Compared to content (i.e. object, scene, composition) diversity, users pay less attention to visual (i.e. color distribution, hue) diversity. Among the four factors in the quality group, aesthetics attracts more concern as a whole, then watermark, size, and aspect ratio.

Users' evaluation criteria distributions are different when search intents vary. Under the "Locate/Learn/Entertain" taxonomy, users select relevance as one of their evaluation criteria in 89% of *Locate* tasks, which are significantly less than that in *Learn* and *Entertain* tasks. Because they need to download images for further use, quality is an important factor only second to relevance. When users just want to get information or kill time with browsing images, the quality of results is less important, while a diverse result list may raise user satisfaction. Under the "Work&Study/Daily-life" taxonomy, a *Work&Study* task pays more attention to the utility (i.e. aspect ratio, size, and watermark) of results, while a *Daily-life* task pays more attention to the diversity and aesthetics. In *What* dimension, *Specific* tasks pay more attention on relevance, while *General* tasks pay more attention on diversity and aesthetics. When users have mental images before the search, they have less concern on diversity and aesthetics.

Relevance is the biggest concern when deciding whether to click on an image result. About 10% to 20% of clicks occur because the image is different from surrounding images, which indicates that being different may attract the user to click. The surrounding text attracts 18% of clicks in *Learn* tasks. About 20% of clicks occur because users are just interested in the image. It happens most frequently in *Entertain* tasks when users have no clear search objective.

5.3 Findings in the user study

In this section, we conduct one-way ANOVA to analyze the effect of search intents in different intent taxonomies on user behavior, evaluation criteria, and reasons for click to address **RQ1** and **RQ2**.

To answer **RQ1**, we conclude that (see Section 5.1 and Table 4): 1) User behavior changes significantly with search intents when classifying tasks by taxonomies of *why* dimension except for the task-related features between *Work&Study* and *Daily-life* tasks. 2) There is no significant difference in most of the users' click behavior when classifying tasks by taxonomies of *what* dimension. 3) The biggest differences between *Specific* and *General* tasks are task length, number of (unique) query terms in query-level. While the biggest differences between *Mental Image* and *Navigation* tasks are the number of terms and Chinese characters in query-level.

To answer **RQ2**, we conclude that (see Section 5.2 and Figure 2): 1) Relevance is the major concern when users perceive satisfaction under all search intents, followed by the quality, diversity, and original website, which are of different importance when search intents vary. 2) Relevance is also the biggest concern when deciding whether to click on an image result. However, the ratio of clicks where users select relevance as the reason for click is lower than the ratio of tasks where users select relevance as the evaluation criterion. Users click on images that they are just interest in (maybe cannot satisfy users' information needs) mostly in *Entertain* tasks.

6 SATISFACTION PREDICTION ACROSS DIFFERENT INTENTS

In this section, we try to predict user satisfaction at the query level with different implicit and explicit signals to answer **RQ3**.

Table 5: Satisfaction prediction performance based on the user behavior and explicit feedback features in terms of AUC score. Results in boldface are the best performance in each row.

Feature Group	*why* dimension					*what* dimension				All tasks
	Locate	Learn	Entertain	Work&Study	Daily-life	Specific	General	Mental Image	Navigation	
Query	**0.630**	0.607	0.530	0.603	0.569	0.615	0.541	0.574	0.582	0.577
Click	0.687	0.614	0.630	0.672	0.669	**0.703**	0.636	0.685	0.668	0.680
Dwell Time	**0.727**	0.594	0.633	0.689	0.636	0.667	0.625	0.664	0.676	0.683
Mouse	0.689	0.633	0.544	0.672	0.670	0.688	0.640	**0.708**	0.695	0.683
Feedback	0.812	0.804	**0.835**	0.805	0.800	0.809	0.821	0.789	0.819	0.813
All Features	**0.864**	0.834	0.831	0.842	0.839	0.842	0.842	0.846	0.858	0.850

Table 6: The list of 35 features extracted from our field study data, which are categorized into five groups.

Feature Group	Description	#
Query	Number of terms/characters in the query text.	2
Click	Number of clicks; First click/last click time; Min/max of click depth (row),	5
Dwell Time	The dwell time on the SERP/Preview page.	2
Mouse	Min/max/median/mean of the distance/speed of mouse movement/scrolling; Moving/scrolling time ratio.	18
Feedback	Min/max/median/mean of the relevance/usefulness score of clicked images from participants' self-annotation.	8

We classify the features into different groups and compare the prediction performance across feature groups and search intents.

The feature groups are listed in Table 6. The former four groups are implicit behavior features which have been introduced in Section 5.1. Query text features reflect the complexity of this search, which can potentially indicate the difficulty level of retrieved results that can satisfy users. The interval between starting a search and the first click action (first click time) reflects the time when they think that the image can satisfy their information needs. We also consider the time users spend (dwell time) both on the SERP and preview page. Features of mouse movements and scrolling such as distance and speed can indicate the searchers' examining behavior. We measure the minimum, maximum, median, and average distance and speed in pixels. The last group is users' explicit feedbacks extracted from the post-questionnaire data, including relevance and usefulness scores of clicked images.

Since the average satisfaction score is between 3 and 4, we map the 5-level satisfaction scale to a 2-level scale (1, 2, 3: USAT; 4, 5: SAT) and treat the prediction task as a binary classification problem. We apply a gradient boosting classifier [26] and perform 5-fold cross validation. The results are shown in Table 5. We can observe that the same feature group performs differently in satisfaction prediction as search intents vary. Query text information performs better in *Locate* tasks, the same to dwell time features. Click features perform better in *Specific* tasks and mouse features perform better in *Mental Image* tasks. Explicit features are more effective than the other features for all types of tasks. It indicates that the relevance and usefulness scores of clicked images strongly correlate with user satisfaction. However, the explicit features are more difficult to obtain. Combining all features together achieves better performance than only using single feature group.

In summary, concerning **RQ3**, we find that different feature groups perform differently in satisfaction prediction as search intents vary. It indicates the importance of considering the search intent to satisfaction prediction in image search.

7 DISCUSSION

Based on the experiment results, we conclude that: i) There is no significant difference in most of users' click behavior signals when classifying tasks according to taxonomies of *what* dimension, while there are significant differences in query text, click, dwell time, and mouse behavior when classifying tasks according to taxonomies of *why* dimension. ii) Relevance is the major concern when perceiving satisfaction under all search intents, followed by the quality, diversity, and original website, which are of different importance when search intents vary. iii) Explicit features such as relevance and usefulness scores of clicked images are more effective than implicit features on the prediction of satisfaction.

We would like to highlight some of the limitations of this work. We did not ask participants to category tasks by themselves, but employed external assessors to make annotations based on their intent descriptions. Even though the annotation results could reach a moderate or higher agreement, it was hard for assessors to judge the intent types in some cases (e.g. participants did not describe their search intents clearly). Meanwhile, we provided options for participants' feedbacks of evaluation criterion and reason for click. Participants tended to select answers from the options rather than input new factors that we did not provide during the field study, which potentially brought in biases on the factors we provided as options. Some other factors may also have effect on users' search experience such as the copyright of images.

8 CONCLUSION AND FUTURE WORK

In this study, we conducted a one-month field study to collect search tasks, intent descriptions, behavior data, and satisfaction scores from participants directly. This dataset can provide a more reliable and realistic view on how search intents are associated with user behavior and satisfaction in image search. We investigate the relationships among search intents, user behavior, and user satisfaction in image search. The experiment results show the importance of considering search intent to better understanding user behavior. Since explicit and implicit signals (e.g. relevance, click, mouse movement) perform differently on satisfaction prediction of tasks with different search intents, it is worth designing evaluation metrics respectively for different search tasks in the future work. Furthermore, we can try to conduct task intent classification models for better understanding search intent and apply it in search satisfaction prediction.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011), National Key Basic Research Program (2015CB358700), and Tsinghua University Tutor Research Fund.

REFERENCES

[1] Paul André, Edward Cutrell, Desney S. Tan, and Greg Smith. 2009. Designing Novel Image Search Interfaces by Understanding Unique Characteristics and Usage. In *Human-Computer Interaction – INTERACT 2009*, Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 340–353.

[2] Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. 2006. The Intention Behind Web Queries. In *String Processing and Information Retrieval*, Fabio Crestani, Paolo Ferragina, and Mark Sanderson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 98–109.

[3] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10.

[4] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information processing & management* 31, 2 (1995), 191–213.

[5] Zhicong Cheng, Bin Gao, and Tie-Yan Liu. 2010. Actively Predicting Diverse Search Intent from User Browsing Behaviors. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/1772690.1772714

[6] Sally Jo Cunningham and David M. Nichols. 2008. How People Find Videos. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '08)*. ACM, New York, NY, USA, 201–210. https://doi.org/10.1145/1378889.1378924

[7] Ovidiu Dan and Brian D. Davison. 2016. Measuring and Predicting Search Engine Users&Rsquo; Satisfaction. *ACM Comput. Surv.* 49, 1, Article 18 (July 2016), 35 pages. https://doi.org/10.1145/2893486

[8] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[9] Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2012. Predicting Web Search Success with Fine-grained Interaction Data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 2050–2054. https://doi.org/10.1145/2396761.2398570

[10] Alan Hanjalic, Christoph Kofler, and Martha Larson. 2012. Intent and Its Discontents: The User at the Wheel of the Online Video Search Engine. In *Proceedings of the 20th ACM International Conference on Multimedia (MM '12)*. ACM, New York, NY, USA, 1239–1248. https://doi.org/10.1145/2393347.2396424

[11] Ahmed Hassan. 2012. A Semi-supervised Approach to Modeling Web Search Satisfaction. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 275–284. https://doi.org/10.1145/2348283.2348323

[12] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: User Behavior As a Predictor of a Successful Search. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/1718487.1718515

[13] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (CIKM '13)*. 2019–2028.

[14] 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 (CHIIR '17)*. ACM, New York, NY, USA, 67–76. https://doi.org/10.1145/3020165.3020188

[15] Peter Ingwersen and Kalervo Järvelin. 2006. *The turn: Integration of information seeking and retrieval in context*. Vol. 18. Springer Science & Business Media.

[16] Bernard J. Jansen. 2008. Searching for digital images on the web. *Journal of Documentation* 64, 1 (2008), 81–101. https://doi.org/10.1108/00220410810844169 arXiv:https://doi.org/10.1108/00220410810844169

[17] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval* 3, 1-2 (2009), 1–224. https://doi.org/10.1561/1500000012

[18] Youngho Kim, Ahmed Hassan Awadallah, Ryen W. White, and Imed Zitouni. 2014. Comparing Client and Server Dwell Time Estimates for Click-Level Satisfaction Prediction. https://www.microsoft.com/en-us/research/publication/comparing-client-server-dwell-time-estimates-click-level-satisfaction-prediction/

[19] Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User Intent in Multimedia Search: A Survey of the State of the Art and Future Challenges. *ACM Comput. Surv.* 49, 2, Article 36 (Aug. 2016), 37 pages.

[20] Christoph Lagger, Mathias Lux, and Oge Marques. 2017. What Makes People Watch Online Videos: An Exploratory Study. *Comput. Entertain.* 15, 2, Article 6 (April 2017), 31 pages. https://doi.org/10.1145/3034706

[21] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. http://www.jstor.org/stable/2529310

[22] Yuelin Li and Nicholas J Belkin. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management* 44, 6 (2008), 1822–1837.

[23] 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 (SIGIR '15)*. ACM, New York, NY, USA, 493–502. https://doi.org/10.1145/2766462.2767721

[24] Mathias Lux, Christoph Kofler, and Oge Marques. 2010. A Classification Scheme for User Intentions in Image Search. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. 3913–3918.

[25] Jiaxin Mao, Yiqun Liu, Ke Zhou, Jian-Yun Nie, Jingtao Song, Min Zhang, Shaoping Ma, Jiashen Sun, and Hengliang Luo. 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)*. ACM, New York, NY, USA, 463–472. https://doi.org/10.1145/2911451.2911507

[26] Llew Mason, Jonathan Baxter, Peter L Bartlett, and Marcus R Frean. 2000. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*. 512–518.

[27] Matthew Mitsui, Jiqun Liu, Nicholas J. Belkin, and Chirag Shah. 2017. Predicting Information Seeking Intentions from Search Behaviors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 1121–1124.

[28] 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.

[29] 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 (CHI '15)*. 985–994.

[30] Hsiao-Tieh Pu. 2005. A comparative analysis of web image and textual queries. *Online Information Review* 29, 5 (2005), 457–467. https://doi.org/10.1108/14684520510628864 arXiv:https://doi.org/10.1108/14684520510628864

[31] Eun Youp Rha, Matthew Mitsui, Nicholas J. Belkin, and Chirag Shah. 2016. Exploring the Relationships Between Search Intentions and Query Reformulations. In *Proceedings of the 79th ASIS&T Annual Meeting: Creating Knowledge, Enhancing Lives Through Information & Technology (ASIST '16)*. American Society for Information Science, Silver Springs, MD, USA, Article 48, 9 pages. http://dl.acm.org/citation.cfm?id=3017447.3017495

[32] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On Fine-Grained Relevance Scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 675–684.

[33] Daniel E. Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. 13–19.

[34] Daniel M Russell, Diane Tang, Melanie Kellar, and Robin Jeffries. 2009. Task behaviors during web search: The difficulty of assigning labels. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE, 1–5.

[35] Geoffrey Underwood and Tom Foulsham. 2006. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *The Quarterly Journal of Experimental Psychology* 59, 11 (2006), 1931–1949. https://doi.org/10.1080/17470210500416342 arXiv:https://doi.org/10.1080/17470210500416342

[36] Pertti Vakkari, Mikko Pennanen, and Sami Serola. 2003. Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing & Management* 39, 3 (2003), 445 – 463. https://doi.org/10.1016/S0306-4573(02)00031-6

[37] Ryen W. White and Diane Kelly. 2006. A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06)*. ACM, New York, NY, USA, 297–306. https://doi.org/10.1145/1183614.1183659

[38] Zhijing Wu, Xiaohui Xie, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. A Study of User Image Search Behavior Based on Log Analysis. In *Information Retrieval*, Jirong Wen, Jianyun Nie, Tong Ruan, Yiqun Liu, and Tieyun Qian (Eds.). Springer International Publishing, Cham, 69–80.

[39] Hong Xie. 1997. Planned and situated aspects in interactive IR: Patterns of user interactive intentions and information seeking strategies. In *Proceedings of the ASIST Annual Meeting*, Vol. 34. 101–10.

[40] 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 Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. 655–663.

[41] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating Examination Behavior of Image Search Users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. 275–284.

[42] 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 *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. 615–624.