Comparing Image Search User Behavior on Lab Study and Field Study Task Settings

Zhijing Wu, Jiaxin Mao, Yiqun Liu*, Min Zhang, and 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
yiqunliu@tsinghua.edu.cn

ABSTRACT

Previous studies have investigated various aspects of user behavior in Web search, such as characteristics of the search session, query reformulation, and click behavior. Most of these studies are based on two types of settings: lab study and field study. Lab study is conducted under a controlled environment, in which it is considered easy to control variables and collect user's search feedback right after each designed search task, but hard to represent real-life information needs. Field study provides practical search behavior data because participants are observed in their own environment with their own search tools. It is of vital importance to understand user behavior in different study settings, because the choice of study settings influences the accuracy of the measurements and the generalizability of the findings. However, little research has been conducted to make quantitative analysis for the consistency and difference on user behavior between these two study strategies. In this paper, we conduct both lab and field studies in Web image search. Then we compare user effort, query strategy, and click behavior in the lab and field study. We find that users put more effort when dealing with search tasks in the field study setting, especially with the *Locate* search tasks proposed by Xie et al. [20]. They spend more time and formulate longer queries in the field study. Users' clicks are more selective in the lab study, while they find the clicked images more useful in the field study. Our findings enhance the understanding of how search task influences users' image search behavior and reveal the difference between lab study and field study, which can help future study design.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**.

KEYWORDS

Search intent; study setting; user behavior

ACM Reference Format:
Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Comparing Image Search User Behavior on Lab Study and Field Study Task Settings. In *Task Intelligence Workshop@The Twelfth ACM International Conference on Web Search and Data Mining (TI@WSDM19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

TI@WSDM19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Search engines have become a primary tool to support searchers' information seeking. Previous research has extensively studied how users interact with search system interface, including the investigation of session characteristics [7], examination behavior [21], click-through, dwell time [15] and fine-grained mouse movement pattern [4]. Interaction models have been widely used for search intent understanding [20], user satisfaction prediction [5], evaluation metrics design [22], and search result ranking [1]. Investigating user's interaction behavior is of vital importance for improving search systems and user experience. Most of related work is based on lab study [4, 21] or practical log analysis [16, 19]. Some work is based on a field study [7, 9]. User behavior in the search process is affected by the study settings. The decision of study methodology should be considered carefully because it influences the accuracy of the measurements and the generalizability of the findings.

In lab studies, search tasks are designed by researchers and the data collection procedure is conducted in the lab environment. Participants can provide explicit feedback such as satisfaction right after each search task. In this setting, the search task is well-defined, but the information needs and user behavior are unnatural. Log-based studies using large-scale search logs are considered to be with a high degree of generalizability because of the large and diverse populations involved. However, researchers need to make assumptions to identify the search tasks. For example, Park et al. [16] divided a user's search activities into separate tasks when the time between consecutive actions exceeds 30 minutes. Collecting explicit feedback from searchers themselves is difficult. Researchers usually employed external assessors to judge whether a task is successful or not [6], which may not accurately reflect users' real experience.

McGrath [14] stated that we can not maximize the two features of generalizability and accuracy only using a single study methodology. Maximizing one feature reduces the other one. Several researchers try to make a trade-off between lab and log-based studies with field studies [7, 9]. A browser plug-in is installed on participants' own laptop to record their natural Web search activities in daily life. They need to make annotations such as search task identification, intent description, and satisfaction feedback for their search logs on a website. This results in a data set that provides a higher accuracy for the task measurements than log-based study and a more natural information seeking behavior than lab study.

*Corresponding author.

To our best knowledge, there is no previous work trying to investigate how user behavior on the lab and field studies differ through quantitative analysis. Meanwhile, users have diverse search intent during the search process. Broder [3] grouped general Web search intent into three categories: *Informational, Transactional, Navigational*. With respect to Web image search, Xie et al. [20] grouped the search intent into *Locate*, *Learn*, and *Entertain*. They also demonstrated that users interact with image search engines in different ways as intent varies. Considering the influence of study setting and search intent on user behavior, we propose our main research question:

RQ: What are the consistency and difference on user behavior between lab and field studies with a certain search intent in Web image research?

To address our research questions, We collected datasets in a lab and a field study respectively. The collected datasets contain user behavior data and search task annotations (such as the usefulness, relevance of search results). Following the search intent taxonomy proposed by Xie et al. [20], we compared the user behavior with certain search intent from three aspects: user effort, query strategy, and click patterns. We find that users put more effort in completing the search tasks in a field study setting, especially with the *Locate* intent. They spent more time and formulated longer queries in the field study. Users clicked on more search results with *Locate* tasks in the field study and *Learn* tasks in the lab study. Users' clicks were more selective in the lab study, while they are easier to feel useful about the clicked images in the field study.

In the rest of this paper, we further review related work and introduce the dataset. Then we report our experiment results and discuss the findings and future work.

2 RELATED WORK

We briefly summarize the related work in three categories: study strategy, search intent taxonomy, and user behavior.

2.1 Study Setting

Considering specific goals of the research, Kelly [10] grouped the strategies for studying information seeking tasks into three categories: exploratory, descriptive, and explanatory studies. The studies can also be characterized according to where the study takes place. Lab studies take place in a controlled lab environment. Naturalistic studies examine information seeking tasks in the same environment as where it occurs. Therefore, we can capture more natural user behavior. The log-based study is a classical example of naturalistic studies. In field studies (another type of naturalistic studies), participants are observed while they are searching in their own environment with their own tools and completing tasks that are motivated by themselves but not the researcher. In this work, we aim to compare the user behavior in the lab and field study.

2.2 Search Intent Taxonomy

Understanding the search intent behind queries is of vital importance. Baeza-Yates et al. [2] concluded that search intent behind queries can be characterized along two dimensions: "what" users are searching for and "why" they search. In image search scenario, Lux et al. [12] proposed a popular intent taxonomy, which categories search intent into navigation, transaction, knowledge orientation, and mental image. A recent study by Xie et al. [20] focused on why people search and grouped image search intent into three categories: *Locate*, *Learn*, and *Entertain*. For *Locate* tasks, users want to find and download images for further use. For *Learn* tasks, users want to learn something, confirm or compare information by browsing images. For *entertain* tasks, users just want to relax and kill time by freely browsing the image search results, where their search behavior is not driven by a clear objective. It is shown that temporal information and mouse movement patterns are useful for distinguishing the search intent.

2.3 Image Search Behavior

In previous work, many features such as query text, session length, browsing depth, and query reformulation patterns are measured to characterize the behavior of users [8, 15]. Recently, Xie et al. [21] used eye-tracking devices to investigate users' examination behavior and found a middle-position bias in users' search patterns. Considering the search intent, Park et al. [16] analyzed a large-scale query log from Yahoo Image Search to investigate user behavior on different query types and identified important behavioral differences across them. In this work, we try to compare the user behavior (which can be logged with Web browser plug-in) of different study strategies.

3 DATASET

To address the research questions, we use two datasets collected from our previous work [18, 23]. In this section, we will make a brief introduction of the datasets. The two datasets are collected through a lab study and a field study respectively. The study procedure is shown in Figure 1. (The details about the datasets and plugin can be found in these two papers [18, 23].)

3.1 Data Collection

To collect search tasks with certain search intent, we follow the most recent intent taxonomy proposed by Xie et al. [20]. According to the criteria of "Is the user's search behavior driven by a clear objective?" (C1) and "Does the user need to download the image for further use after the search process?" (C2), search intent is divided into three groups:

- **Locate**: The user is looking to download something for further use. Example: I want to change the desktop background of this computer. The content of background should contain the forest and blue sky. (C1 = 1, C2 = 1)
- **Learn**: The user is looking to discover something or learn about a topic, confirm or compare information by browsing images. They can obtain, check or compare information by examining images in result pages only. Example: I bought a white linen t-shirt yesterday. I want to see which pants and shoes can match it. (C1 = 1, C2 = 0)
- **Entertain**: The user just wants to browse images for fun in order to kill time. Example: I have nothing to do and just want to browse some posters or photos of my favorite stars. (C1 = 0, C2 = 0)

3.1.1 Lab Study. Following the search intent taxonomy introduced above, we design 12 search tasks (There are 4 search tasks in each

We list the 12 search tasks in Appendix A.

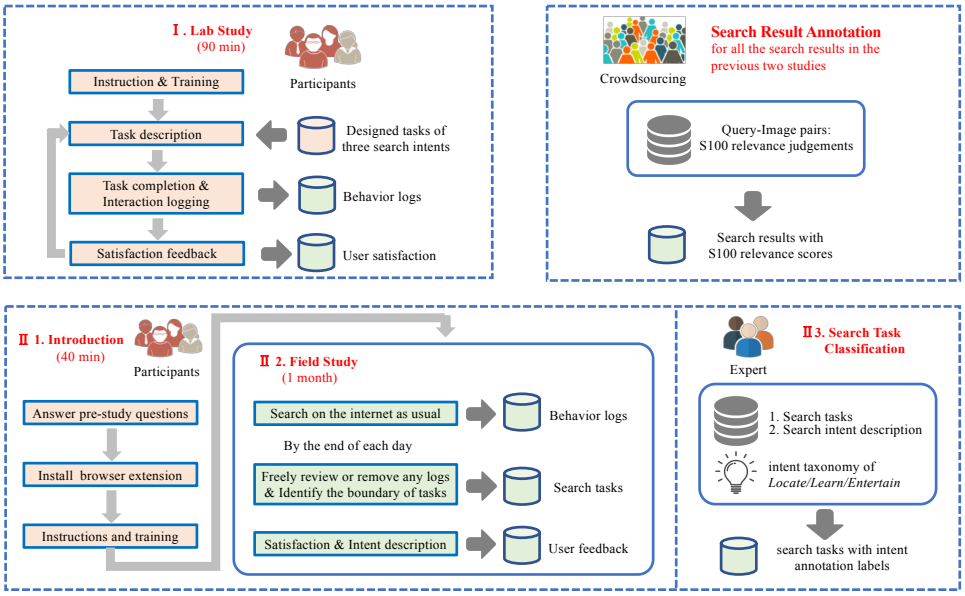

Figure 1: The procedure of lab study (part I) and field study (part II). S100 relevance judgments of these two studies are collected through crowdsourcing.

intent category) for the lab study. Participants are invited to our lab to finish all the tasks with computers provided by us after completing the training process. For each task, a detailed task description is provided. Participants can use the search engine as usual to complete the search tasks, during which their search behavior including inputting queries, clicking on results, mouse scrolling and movement is recorded automatically. After that, they are required to give 5-point scaled satisfaction scores [11] (1: unsatisfied, 2: slight satisfied, 3: fair satisfied, 4: substantial satisfied, 5: very satisfied) for each query and task. Meanwhile, they are required to give 4-point scaled usefulness scores [13] (1: not useful at all, 2: somewhat useful, 3: fairly useful, 4: very useful) for clicked images.

3.1.2 Field Study. In the field study, we develop a Web browser plug-in and an annotation website for recording the search behavior and collecting search feedback. Before the field study, participants are invited to the lab with their own laptops. They are instructed to install the Web browser plug-in and use the annotation website. After that, they can use their laptops as usual and we will record their image search activities in the following month. By the end of each day, participants are required to log into the annotation website to examine their search logs. They are allowed to remove any logs that they do not want to share with us. Meanwhile, they need to identify the queries that were submitted to complete the same search task, and give a search intent description for the task. Finally, they are required to give a 5-point scaled satisfaction score for each query and task as well as a 4-point scaled usefulness score for each clicked image.

After the field study, external experts are recruited to annotate the intent category (i.e. *Locate*, *Learn*, or *Entertain*) for each search task. Experts are firstly gathered to discuss the classification criterion. Then the participant's intent description and the query list for each task are shown for the judgments. Each task is annotated by three experts. We use the majority vote of three annotations as the category label of the search task.

Table 1: Description of the source of search tasks, intent categories, satisfaction and relevance.

	Study Strategies	
	Lab Study	**Field Study**
Task defining	Researcher	User
Task classification	Researcher	Expert
S5 Satisfaction	User	User
S4 Usefulness	User	User
S100 relevance	Crowdsourcing	Crowdsourcing

With the search tasks collected through the lab and field studies, we further collect the fine-grained relevance judgments for all image results through crowdsourcing. We follow the method proposed by Roitero et al. [17] to get relevance judgments. The S100 relevance judgments range from 0 to 100 and have been verified to give annotators more flexibility than traditional coarse-grained scales. We provide the search query and image result to collect S100 relevance judgments for each query-image pair from 5 different annotators and consider the arithmetic mean of 5 scores as the aggregated relevance score.

Table 1 shows the summary of sources of search tasks, intent categories, satisfaction, usefulness, and relevance. In the field study, the task boundary is identified by user themselves. Users provide satisfaction and usefulness feedback right after search process in the lab study, while they make annotations by the end of each day in the field study.

3.2 Collected Data

We recruit 36 participants (14 females and 22 males) in the lab study and 50 participants (23 females and 27 males) in the field study. All the participants report that they are familiar with the search

Table 2: User behavior features and their descriptions.

	Feature	Description
Effort	TaskLength	Number of queries submitted to complete the task.
	TaskDwellTime	Dwell time of the task.
Query Strategy	AvgQCharacter	Average number of characters in queries of the task.
	AvgQTerm	Average number of terms in queries of the task.
	UQTerm	Number of unique query terms in the task.
	AvgRQTerm	Average number of repetitive query terms between two consecutive queries in the task.
Click	#Click	Number of clicks in the task.
	TimeToFirstClick	Average of time delta between the start of a search session and the first click.
	TimeToLastClick	Average of time delta between the start of a search session and the last click.

Table 3: Average user behavior statistics for the comparison between lab study and field study tasks with certain search intent.

		Locate		Learn		Entertain		All Tasks	
		Lab	Field	Lab	Field	Lab	Field	Lab	Field
Effort	TaskLength	3.89	**4.25**	2.98	3.45	2.11	3.35	2.95	3.68
	TaskDwellTime(min)	**5.82**	**5.80**	3.67	3.73	1.75	2.38	3.65	3.95
Query Strategy	AvgQCharacter	5.62	**6.28**	5.77	6.07	4.56	5.44	5.29	5.93
	AvgQTerm	2.68	2.53	**3.16**	2.25	2.32	2.21	2.71	2.33
	UQTerm	**6.06**	5.78	5.37	4.85	4.04	4.73	5.11	5.11
	AvgRQTerm	0.65	0.79	**0.97**	0.65	0.33	0.58	0.64	0.67
Click	#Click	4.58	5.98	**6.14**	4.96	5.71	3.54	5.50	4.82
	TimeToFirstClick(s)	22.05	**25.69**	15.33	20.37	9.62	12.27	15.85	19.22
	TimeToLastClick(s)	53.32	**57.31**	48.09	41.82	35.68	21.75	46.04	39.66

Table 4: Statistics of the search data collected through lab study and field study.

		Locate	Learn	Entertain	All
Lab Study	#Tasks	118	125	136	379
	#Queries	459	373	287	1,119
Field Study	#Tasks	179	192	184	555
	#Queries	760	663	617	2,040

engines and have experience in performing Web image search tasks. The statistics of the search data are shown in Table 4. We collect 379 lab study tasks and 555 field study tasks. There are 32.3% *Locate*, 34.6% *Learn*, 33.1% *Entertain* tasks in the field study data, which indicates the real user information needs distribution.

4 RESULTS

With the image search collected in the two studies, we respectively analyze users' behavior in the lab study and field study to address our research question. Table 2 lists the features that we focus in this work. These features cover three aspects: user effort, query strategy, and click patterns. The statistical results are shown in Table 3.

4.1 User Effort

We use the number of queries and the total dwell time in completing a single task to measure user's effort. As shown in Table 3, *Locate* task is the most difficult one to finish (for which users submit more queries and spend more time), followed by the *Locate* and *Entertain* tasks. Users submit more queries to finish one task with all types of search intents in the field study. They spend more time with the *Entertain* tasks in the field study than in the lab study. It indicates that when users conduct search activities in the controlled lab environment, they tend to finish the search tasks as soon as possible. Especially when they do not have a clear objective (with the *Entertain* intent), they spend less time on browsing images in the lab study (1.75 minutes on average, which is less than 2.38 minutes in the field study).

4.2 Query Strategy

We analyze how user formulate their queries in completing the search tasks. We calculate the average number of Chinese characters and terms in queries of a search task. We segment the query text (Chinese) into words and remove the stop words. The "query term" refers to words in the processed query text. We find that the average number of Chinese characters in the lab study is smaller than that in the field study, while the average number of query terms in the lab study is larger than that in the field study. It indicates the length of each query terms are longer in the field study. In the task level, we calculate the number of unique query terms in each task. We find that there are more unique query terms in the field study than in the lab study with *Entertain* intent (4.73 versus 4.04). We calculate the number of repetitive query terms with the last query in the same task (AvgRQTerm). We find that there are less repetition with the *Entertain* intent. There are more repetition between two consecutive queries in the field study compared to the lab study except for the *Learn* search tasks. It may because that we provide more detailed task descriptions for the *Locate* tasks and give some examples of search aspects (e.g. Chinese style and Simple European

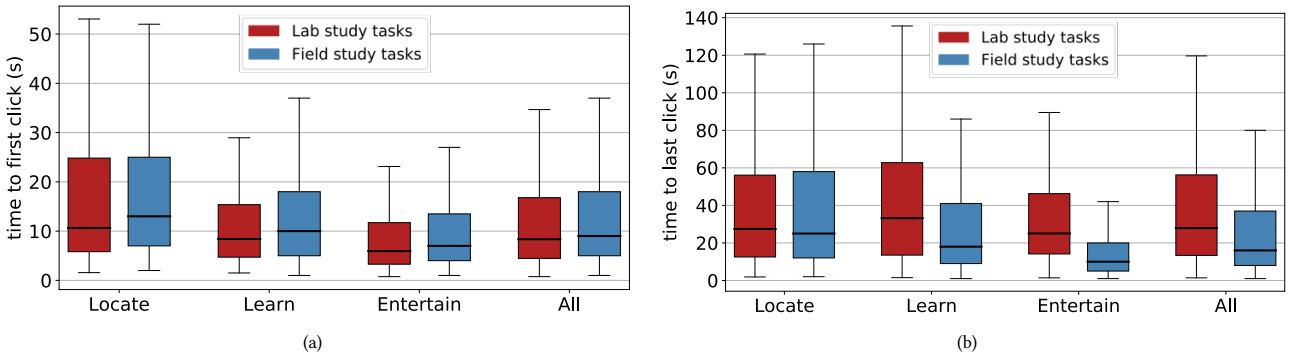

Figure 2: Distribution of the time to (a) first click; (b) last click.

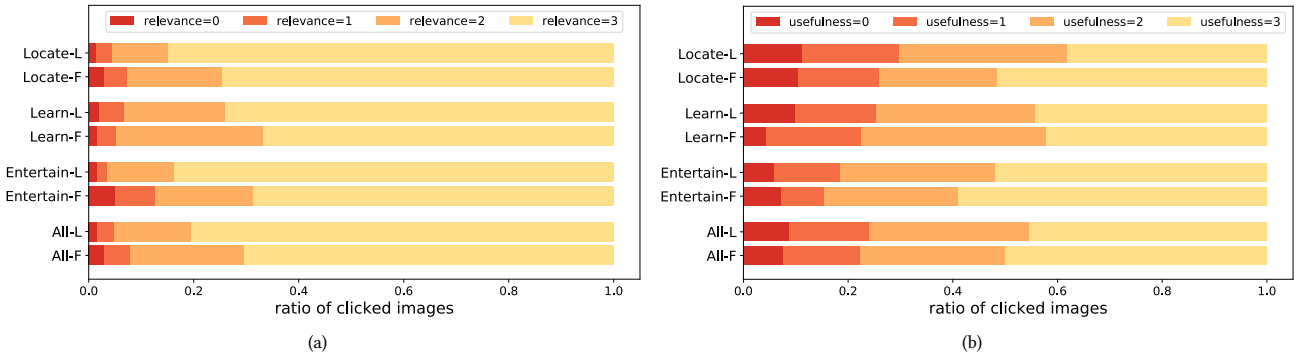

Figure 3: Distribution of the (a) relevance score; (b) usefulness score of clicked images. ("Locate-L" refers to the *Locate* tasks with lab study setting and "Locate-F" refers to the *Locate* tasks with field study setting.)

style for the sixth task in Table 5). Users formulate their search queries with focus on these aspects.

4.3 Click Behavior

Users click on more image results in the field study tasks with *Locate* intent, while they click on less image results in the field study tasks with *Learn* and *Entertain* intent. We plot the distribution of the time to first and last click in Figure 2. Users' first click on an image occurs earlier in the lab study than in the field study during a search session, which is consistent with the finding in Section 4.1 that user tend to finish the search tasks as soon as possible in the lab study.

With the S100 relevance judgments on all image results, we map them to 4-point scale relevance judgments (0: [0,25); 1: [25, 50); 2: [50, 75); 3: [75, 100]) and plot the distribution of 4-point scale relevance (annotated through crowdsourcing) and usefulness (annotated by user themselves) scores of clicked images in Figure 3. The relevance level of clicked images in the lab study is higher than that in the field study, which indicates that users' clicks are more selective in the lab study. However, one image result is relevant to the query does not mean that it is useful for the users. Users find the clicked images more useful in the field study. It may because that in the lab study, we provide the detailed task descriptions for users. They have more strict criterion on judging whether an image is relevant to the query.

4.4 Summary of Findings

In this section, we analyze the effect of study settings and search intents on user effort, query strategy, and click behavior to address our research question. We conclude that: 1) Users put more effort when dealing with search tasks in the field study setting, especially with the *Locate* search tasks. 2) Users formulate longer queries in the field study and there are more repetition between two consecutive queries except for the *Learn* search tasks. 3) In the field study, users clicks on more search results with *Locate* intent, while with *Learn* intent in the lab study. 4) Users' clicks are more selective in the lab study (the relevance scores of clicked images are higher in the lab study than that in the field study), while they find the clicked images more useful in the field study.

5 CONCLUSION AND FUTURE WORK

In this work, we use the dataset from the lab and field studies in Web image search scenario to analyze user behavior with certain intent. The experiment results shows the differences on user behavior between the lab study and field study task settings. Users put more effort when dealing with search tasks in the field study setting. They formulate longer queries and spend more time in the field study. Users' clicks are more selective in the lab study, while they find the clicked images more useful in the field study. These results show that some findings in the lab study may not generalize to the more natural field study setting. Future research on how search intent affects search behavior should acknowledge such differences. In the future work, we would like to compare the process of satisfaction perception as well as conduct the comparison studies in more search scenarios such as video search and mobile search, in which the user interfaces are different with desktop image search.

Table 5: Search tasks for lab study.

Intent	Task ID	Task Description
Locate	1	Imagine you are invited to design a poster for a dancing party which will be held this weekend. Please find some related images.
	2	Imagine you want to find useful pictures for a short news report of 2016 US presidential election.
	3	Imagine you want to make slides about Harry Potter. You need some posters of Harry potter film.
	4	Please change the desktop background of this computer, the content of background should contain the forest and blue sky.
Learn	5	Imagine you just receive a job offer in New York City. You want to know more about this City (e.g. streets, landscapes, buildings).
	6	Imagine you prepare to renovate a new house. You would like to compare different decoration styles (e.g. Chinese style, Simple European style).
	7	Imagine you bought a white linen t-shirt yesterday, you want to see which pants and shoes can match it.
	8	Imagine you saw a beautiful flower on the way to school. The flower has white petal and yellow stamen, you want to find out its name (We already provide a picture about this flower, please check it before searching).
Entertain	9	Imagine you want to browse some posters or photos of your favorite stars.
	10	Imagine you want to search for some humorous pictures to relax yourself.
	11	Imagine you want to browse some posters or pictures of your favorite movies.
	12	Imagine you want to browse some pictures of your favorite cartoons.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and The National Key Research and Development Program of China (2018YFC0831700).

A SEARCH TASKS OF LAB STUDY

The search tasks used in lab study is listed in Table 5.

REFERENCES

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving Web Search Ranking by Incorporating User Behavior Information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. 19–26.

[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] 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)*. 2050–2054.

[5] 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)*. 275–284.

[6] 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)*. 221–230.

[7] 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)*. 67–76.

[8] Bernard J. Jansen. 2008. Searching for digital images on the web. *Journal of Documentation* 64, 1 (2008), 81–101.

[9] Melanie Kellar, Carolyn Watters, and Michael Shepherd. 2007. A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 999–1018.

[10] 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

[11] 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

[12] 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.

[13] 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)*. 463–472.

[14] JOSEPH E. McGrath. 1995. METHODOLOGY MATTERS: DOING RESEARCH IN THE BEHAVIORAL and SOCIAL SCIENCES. In *Readings in HumanâĂŞComputer Interaction*, RONALD M. BAECKER, JONATHAN GRUDIN, WILLIAM A.S. BUXTON, and SAUL GREENBERG (Eds.). Morgan Kaufmann, 152 – 169.

[15] 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.

[16] 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.

[17] 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.

[18] Zhijing Wu, Yiqun Liu, Qianfan Zhang, Kailu Wu, Min Zhang, and Shaoping Ma. 2019. The influence of image search intents on user behavior and satisfaction. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*.

[19] 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.

[20] 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)*. ACM, New York, NY, USA, 655–663. https://doi.org/10.1145/3159652.3159686

[21] 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.

[22] 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 (SIGIR '17)*. 425–434.

[23] 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 & Development in Information Retrieval (SIGIR '18)*. 615–624.