Query/Task Satisfaction and Grid-based Evaluation Metrics Under Different Image Search Intents

Kosetsu Tsukuda and Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{k.tsukuda,m.goto}@aist.go.jp

ABSTRACT

People use web image search with various search intents: from serious demands for work to just passing time by browsing images of a favorite actor. Such a diversity of intents can influence user satisfaction and evaluation metrics, both of which are important factors for providing a better image search environment. In this paper, we investigate this influence by using a publicly available one-month field study dataset. With respect to satisfaction, we take into consideration both query-level and task-level satisfaction provided by search users. Regarding the evaluation metrics, we use grid-based evaluation metrics that incorporate user behavior specific to image search. The results of our analysis indicate that both query/task satisfaction and grid-based evaluation metrics are influenced by the image search intent. Based on the results, we show possibilities to support users' search processes according to their search intents. We also discuss that there is still room for improvement in evaluation metrics through the development of intent-aware evaluation metrics in image search.

ACM Reference Format:
Kosetsu Tsukuda and Masataka Goto. 2020. Query/Task Satisfaction and Grid-based Evaluation Metrics Under Different Image Search Intents. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401295

1 INTRODUCTION

Web image search is used for various search intents. For example, some users want to download images for further use after searching them online, while others want to learn, confirm, or compare specific information by browsing images [11]. Under such intents, a user addresses a specific image search task (e.g., a task to learn about the atmosphere of the city of Xi'an before visiting there under the learning intent). The search task process influences the user's satisfaction: if she successfully completes the task, the degree of satisfaction is high; if she cannot find the desired images even though she submitted many queries, the degree is low. Naturally, the concepts of intent, task, and satisfaction are closely related.

Regarding satisfaction, we can think about satisfaction at the query level and the task level, where a task consists of one or more queries submitted by the user. With respect to the task satisfaction, users may easily achieve high degree of satisfaction when performing search tasks with certain intents, while it may be difficult with some other intents. It has been reported that users' intents can be effectively predicted from their behavior in the early stages of the search task [11]. If the intent thus predicted shows low satisfaction, it would be helpful to support the users' image search at this point. Nevertheless, characteristics of task satisfaction for different search intents have not been studied well. As for the query satisfaction, Wu *et al.* [10] proposed a method for predicting user satisfaction at the query level in accordance with the search intent. The method is based on their observation that user behavior during the search process varies depending on whether the queries are satisfied or not. Although they have showed that satisfaction can be predicted with high accuracy regardless of the intent, not enough effort has been made towards understanding the relationship between query satisfaction and search intent. For example, it is not clear whether a user stops searching for images after having submitted a satisfied query in each search intent. Moreover, because users usually submit several queries to achieve the task goal, it is natural to think that the query satisfaction influences the task satisfaction. However, to the best of our knowledge, there are no studies in the image search field that investigate the relationship between the query-level and the task-level satisfaction. This relationship would allow us to understand user behavior at a deeper level. Considering all of the above, we address the following research question:

RQ1 *What are the characteristics of the query and the task satisfaction and what is the relationship between them under different image search intents?*

By answering this question, we aim to reveal an appropriate approach to support the users according to their image search intent.

User satisfaction also plays an important role in evaluation metrics because the effectiveness of metrics is typically evaluated in terms of the correlation with user satisfaction [13]. In image search, as in general web search, developing better evaluation metrics is an important research topic because it enables the search engine developers to accurately evaluate the engine's performance and revise the image-ranking algorithms [12]. Recently, Xie *et al.* [13] have proposed grid-based evaluation metrics for image search. Because image search results are typically placed on a grid-based panel, user behavior on the search engine result page (SERP) of image search is different from that on the SERP of general web search. Xie *et al.* [13] found three major types of behavior specific to image search, namely, "Middle bias," "Slower decay," and "Row skipping" (see Section 4 for details). Their proposed metrics reflect such behavior. Xie *et al.* [13] showed the effectiveness of the proposed metrics against the metrics for general web search such as Rank-Biased Precision (RBP) [8] and Discounted Cumulative Gain

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401295

Table 1: Dataset statistics.

	why dimension					*what* dimension			
	Purpose ∗∗ ★★			Situation ∗∗ ★★		Concreteness ★★		Content	
	Locate	Learn	Entertain	Work&Study	Daily-life	Specific	General	Mental Image	Navigation
Number of tasks	151	153	143	167	280	201	246	133	314
Avg task satisfaction	3.77	3.83	4.20	3.69	4.06	3.85	4.01	3.90	3.94
Avg query satisfaction	3.24	3.59	4.09	3.20	3.88	3.48	3.70	3.69	3.58

∗∗ (★★) denotes the statistical difference at $p < 0.01$ among different intents of a taxonomy regarding the average of task satisfaction (the average of query satisfaction).

(DCG) [5]. However, they do not investigate the performance of the metrics across search intents. This motivates us the following research question:

RQ2 *How do image search intents affect the performance of the grid-based evaluation metrics?*

Answering this question is beneficial for improving the evaluation metrics. The study closest to our motivation is the one by Zhang *et al.* [14]. Our study is different from theirs in that: (1) they consider only three search intents (*Locate*, *Learn*, and *Entertain* [11]), while we use nine intents including those three; and (2) they use only the metrics for general web search, while we use the metrics for image search in addition to the metrics for general web search.

To answer these research questions, we use a publicly available dataset developed through a field study [10]. The data collected from a field study has advantages over the log data from a commercial search engine [9] and from a lab study [14] because the field-study data can provide more accurate annotations and more practical search behavior data [10]. The reusable insights obtained from the results of our analysis can be summarized as follows:

(1) Users who browse images to pass time or search images for daily life tend to achieve high query/task satisfaction. In contrast, users who have more demanding intents tend to have low query/task satisfaction. Helping such users to submit their first query in a task is one possible way to increase their satisfaction because they struggle more to get satisfied results by the first query. In addition, for certain intents such as learning something and looking for general information rather than specific one, it is beneficial to support users' search processes even after they found a desired image because for these intents, submitting many satisfied queries contributes to increase the task satisfaction.
(2) When users want to learn something or find images for daily life, or when users know how the image content looks like before submitting a query, it is effective to incorporate the behavior specific to the image search into the evaluation metrics. For other intents, there is still room for improvement in evaluation metrics by, for example, developing intent-aware metrics.

2 DATASET

2.1 Original Dataset

In this study, we use a publicly available field study dataset provided by Wu *et al.* [10]. The field study logged participants' daily image search activities by using a web browser plugin. Participants were required to identify the queries belonging to the same search task. They were also asked to provide an explicit 5-level satisfaction feedback (1: least, 5: most) for each query and each task. Through a one-month field study, 555 tasks, 2,040 queries, and 270,315 image results were collected from 50 participants. In addition to the feedback from the participants, for each query-image pair, at least five assessors recruited via crowdsourcing labeled the relevance scores that ranged from 0 to 100. Moreover, 12 assessors from a commercial search engine company annotated search intents for each task. In this process, four search intent taxonomies [4, 6, 7, 11] were used. Among them, two taxonomies represent "why users search for images" (*why* dimension). One taxonomy consists of three intents while the other consists of two intents:

Locate, Learn, Entertain [11]. *Locate* intent indicates that the users want to download images for further use after the search process. With *Learn* intent, users want to learn, confirm, or compare information by browsing images. Users with *Entertain* intent want to relax and pass time by freely browsing the image search results (e.g., browsing photos of the user's favorite actor). We refer to this taxonomy as *Purpose*.

Work&Study, Daily-life [4]. Users who have *Work&Study* intent want to find images used for their work or study, while search activities related to neither work nor study have *Daily-life* intent. We call this taxonomy *Situation*.

The remaining two taxonomies represent "what kind of images users search for" (*what* dimension), each of which consists of two intents:

General, Specific [6]. Under *General* intent, users want to find general information about a subject, while under *Specific* intent, users have explicit or concrete goals. We refer to this taxonomy as *Concreteness*.

Mental Image, Navigation [7]. Users who have *Mental Image* intent know how the image content looks like and want to find images that match to their mental image. Under *Navigation* intent, users know about the existence of the image, but its content is unknown. We call this taxonomy *Content*.

Following the definitions of each intent described above, assessors assigned one intent from each taxonomy to a task (i.e., four intents were assigned to each task).

2.2 Filtered Dataset

In the original dataset, some query-image pairs have invalid relevance scores; therefore, queries that include one or more images with such invalid scores are removed. To compute the evaluation metrics in Section 4, queries that have less than five rows in the grid-based search results are removed [13]. We then remove users who submitted fewer than three tasks. The filtered dataset consists of 29 users, 447 tasks, 1,758 queries, and 205,306 images. Distribution of the search task intents among 447 tasks is shown in Table 1. In the same table, we also show the average task/query satisfaction for each intent. Regarding the relevance score of each query-image pair, we follow Xie *et al.* [13] and use the average of scores annotated by assessors to each pair.

Table 2: The average number of submitted queries in a task under various conditions.

	why dimension					*what* dimension			
	Purpose ∗			Situation ∗		Concreteness ★ ‡ ‡		Content	
Condition	Locate	Learn	Entertain	Work&Study	Daily-life	Specific	General	Mental Image	Navigation
Before the first satisfied query	0.658	0.363	0.206	0.598	0.302	0.337	0.477	0.236	0.468
Before the first satisfied query in a difficult task	2.32	1.80	1.42	2.24	1.70	1.49	2.55	1.39	2.10
After the first satisfied query	3.29	2.65	2.90	3.24	2.79	1.97	4.04	2.49	3.12

∗/★ denotes the statistical difference at $p < 0.05$ among different intents of a taxonomy in the first/second condition.
‡‡ denotes the statistical difference at $p < 0.01$ among different intents of a taxonomy in the third condition.

Table 3: Pearson's Correlation between task satisfaction and average/maximum of query satisfaction in each task.

	why dimension					*what* dimension			
	Purpose			Situation		Concreteness		Content	
	Locate	Learn	Entertain	Work&Study	Daily-life	Specific	General	Mental Image	Navigation
Avg	0.822	0.810 †	0.763	0.814	0.795 †	0.815	0.799 †	0.853 †	0.786
Max	0.825	0.741	0.742	0.824	0.735	0.818	0.725	0.804	0.773

† denotes the statistical difference at $p < 0.05$. All correlations are significant at $p < 0.001$.

3 QUERY/TASK SATISFACTION

In this section, we answer **RQ1**. Table 1 shows a significant difference among different intents of each taxonomy; one-way ANOVA is used for *Purpose* taxonomy, while t-test is used for the other three taxonomies. With respect to the task satisfaction, it can be observed that only taxonomies in *why* dimension show a significant difference between intents (in the *Purpose* taxonomy, t-test results indicate that *Entertain* has statistically higher satisfaction than *Locate* and *Learn*). It is interesting that in *what* dimension, even if users have no specific goal (*General*) or do not know how the image content looks like (*Navigation*), they finally achieve high enough satisfaction. In *why* dimension, the results where *Entertain* and *Daily-life* have higher satisfaction match with our intuition because the required images in these intents are not so demanding compared to the other intent(s) in the same taxonomy. Regarding query satisfaction, in *why* dimension, as in the case of task satisfaction, *Entertain* and *Daily-life* show statistically higher satisfaction. In *what* dimension, significant difference is found in *Concreteness* taxonomy. We can therefore say that the intents with lower satisfaction are the difficult intents. Because user intents can be effectively predicted from the user behavior in the early stage of a search session [11], it would be beneficial to propose architectures to support the search process of those users who have such difficult intents. However, how can we support users based on their search intents? To answer this question, below, we further investigate the relationships between the task satisfaction and the query satisfaction under different search intents.

We first analyze the average number of submitted queries in a task under various conditions. In this analysis, we use tasks that include at least one satisfied query, where the satisfied query is defined as the query whose satisfaction level is 4 or 5 [10]. Among 447 tasks, 375 tasks meet this definition. In the first condition, we compare the average number of submitted queries in a task before the first satisfied query. The comparison results are shown in Table 2. Because the average values in all intents are statistically lower than 1 at $p < 0.05$, users usually can obtain satisfied search results by the first query. However, the values of *Locate* and *Work&Study* are statistically higher than those of *Entertain* and *Daily-life*, respectively. These results indicate that one factor of the task difficulty in intents of *why* dimension is the difficulty of getting satisfied search results by the first query. Therefore, to increase the task satisfaction of users who have *Locate* or *Work&Study* intents, it would be helpful, for example, to suggest more related queries when they submit their first query.

In the second condition, we focus on difficult tasks in which the first query is not satisfied (i.e., satisfaction level is lower then 4) and count the average number of submitted queries in a task before the first satisfied query. In Table 2, we can see that a significant difference is observed only in *Concreteness* taxonomy. Remind that in *Concreteness* taxonomy, no significant difference was observed between *Specific* and *General* at the task level satisfaction (Table 1). These results indicate that users who have *General* intents struggle more to revise the first unsatisfied query than those who have *Specific* intents. Hence, it would be possible to increase the task satisfaction under *General* intents by, for example, suggesting subtasks estimated from the first query [3].

In the third condition, we compare the average number of submitted queries in a task after the first satisfied query. The results are shown in the bottom row of Table 2. In this condition, too, significant difference is observed in *Concreteness* taxonomy. The number of submitted queries is statistically higher than 3 at $p < 0.05$ in *General* and is statistically higher than 2 in the remaining six intents, except for *Specific* and *Mental Image*. This means that in image search, users tend to submit many queries even after they find the desired images; but does that contribute to improve their task satisfaction? To reveal this, we investigate the correlation between task satisfaction and average/maximum of query satisfaction in each task, as indicated by Avg/Max in Table 3. In *Learn*, *Daily-life*, *General*, and *Mental Image* intents, the values of Avg are statistically higher than the values of Max. In these intents, submitting many high-satisfaction queries contributes to increase task satisfaction more than submitting one fully satisfied query. Therefore, it makes sense to support users even after they submitted a satisfied query. On the other hand, in the remaining five intents where no significant difference is observed between Avg and Max in Table 3, when the system predicts a user submitted a satisfied query, it may contribute to increase the task satisfaction to suggest that the user should not search for images anymore in the task.

Table 4: Pearson's Correlation between evaluation metrics and query satisfaction.

	why dimension					*what* dimension			
	Purpose			Situation		Concreteness		Content	
Metric	Locate	Learn	Entertain	Work&Study	Daily-life	Specific	General	Mental Image	Navigation
RBP	0.304	0.401	0.360	0.299	0.334	0.389	0.272	0.377	0.314
RBP-MB	0.304	0.429 ††	0.379	0.299	0.372 ††	0.399	0.286	0.411 ††	0.319
RBP-SD	0.309	0.433	0.381	0.294	0.390 ††	0.399	0.290	0.420	0.320
RBP-RS	0.302	0.400	0.362	0.302	0.337	0.387	0.274	0.371	0.318

†† denotes the statistical difference at $p < 0.01$ with RBP. All correlations are significant at $p < 0.001$.

4 GRID-BASED EVALUATION METRICS

Evaluation Metrics. Evaluation metrics for general web search such as RBP [8] and DCG [5] can be generalized as a function of gain and stopping probability [1]. By revising the stopping probability, Xie *et al.* [13] proposed grid-based evaluation metrics for image search. To be more specific, they incorporated user examination behavior on the SERP specific to image search. Three characteristics of such behavior are "Middle bias," "Slower decay," and "Row skipping." "Middle bias" indicates that users tend to pay more attention to images in the middle horizontal position on the image SERP than to those in the leftmost or rightmost positions. "Slower decay" means user attention on the image SERP decays more slowly than on a general web SERP (i.e., image search has deeper browsing depths on the SERP). "Row skipping" means that users may skip particular rows on the image SERP and jump to the results at some distance. Note that Xie *et al.* [13] do not incorporate all of the behavior in one evaluation metric; they propose three evaluation metrics where each metric incorporates one of the three characteristics of the behavior. More details about the way to incorporate the behavior in the stopping probability can be found in Xie *et al.* [13].

Experimental Setup. To answer **RQ2**, we evaluate whether the effective behavior varies under different search intents. Following Xie *et al.* [13], given a search intent and an evaluation metric, we measure the effectiveness of the metric by Pearson's Correlation between metric scores and user satisfaction for queries submitted in all the tasks of the intent. As mentioned above, the grid-based evaluation metrics require a baseline evaluation metric for general web search. We use RBP as a baseline model because it has been reported that RBP-based metrics are more effective than other baseline models such as Expected Reciprocal Rank (ERR) [2] and DCG [13]. Grid-based evaluation metrics that incorporate "Middle bias," "Slower decay," and "Row skipping" are indicated by "RBP-MB," "RBP-SD," and "RBP-RS," respectively. Because all of the metrics have hyper-parameters, we tune them by using a grid-search as was done by Xie *et al.* [13]. In RBP-SD, we assume the "Row skipping" behavior starts from the second row [13].

Results. Table 4 shows the results. In the study by Xie *et al.* [13], the metrics were evaluated without considering search intents. The authors showed that their proposed gird-based evaluation metrics are more effective than RBP; they did not, however, report the significant difference. As can be seen in Table 4, when intents are taken into consideration, RBP-MB is most effective because it statistically outperforms RBP in three intents (*Learn*, *Daily-life*, and *Mental Image*). RBP-SD is also effective in *Daily-life* intent (no significant difference is observed between RBP-MB and RBP-SD). In six intents except for *Learn*, *Daily-life*, and *Mental Image*, the impact of the behavior specific to the image search is not significant. From these results, we can conclude that the effective behavior does vary under different intents, and that the grid-based evaluation metrics can help RBP to achieve better correlation with user satisfaction when users have *Learn*, *Daily-life*, or *Mental Image* intents. In the remaining six intents, given that user's image search behavior varies largely depending on the intents [10], it is promising to propose intent-aware evaluation metrics by incorporating behaviors specific to each intent.

5 CONCLUSION

In this paper, we investigated the influence of image search intent on query/task satisfaction and grid-based evaluation metrics by using a field study dataset. Analysis results of user satisfaction indicated various possibilities to support users according to their intent. Regarding evaluation metrics, we showed that it is effective to incorporate the behavior specific to image search into an evaluation metric when users have *Learn*, *Daily-life*, or *Mental Image* intent. As our future work, we plan to consider a combination of intents (e.g., users who have *Learn* and *Specific* intents). Analyzing the influence of such combined intents will enable a deeper understanding of user satisfaction and evaluation metrics in image search.

Acknowledgments. This work was supported in part by JST ACCEL Grant Number JPMJAC1602, Japan.

REFERENCES

[1] L. Azzopardi, P. Thomas, and N. Craswell. Measuring the utility of search engine result pages: An information foraging based measure. SIGIR, page 605–614, 2018.
[2] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. CIKM, page 621–630, 2009.
[3] A. Hassan Awadallah, R. W. White, P. Pantel, S. Dumais, and Y.-M. Wang. Supporting complex search tasks. CIKM, page 829–838, 2014.
[4] P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Springer Science & Business Media, 2005.
[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.
[6] Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management*, 44(6):1822–1837, 2008.
[7] M. Lux, C. Kofler, and O. Marques. A classification scheme for user intentions in image search. CHI EA, page 3913–3918, 2010.
[8] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1), 2008.
[9] J. Y. Park, N. O'Hare, R. Schifanella, A. Jaimes, and C.-W. Chung. A large-scale study of user image search behavior on the web. CHI, page 985–994, 2015.
[10] Z. Wu, Y. Liu, Q. Zhang, K. Wu, M. Zhang, and S. Ma. The influence of image search intents on user behavior and satisfaction. WSDM, page 645–653, 2019.
[11] X. Xie, Y. Liu, M. de Rijke, J. He, M. Zhang, and S. Ma. Why people search for images using web search engines. WSDM, page 655–663, 2018.
[12] X. Xie, J. Mao, Y. Liu, M. de Rijke, Q. Ai, Y. Huang, M. Zhang, and S. Ma. Improving web image search with contextual information. CIKM, page 1683–1692, 2019.
[13] X. Xie, J. Mao, Y. Liu, M. de Rijke, Y. Shao, Z. Ye, M. Zhang, and S. Ma. Grid-based evaluation metrics for web image search. WWW, page 2103–2114, 2019.
[14] F. Zhang, K. Zhou, Y. Shao, C. Luo, M. Zhang, and S. Ma. How well do offline and online evaluation metrics measure user satisfaction in web image search? SIGIR, page 615–624, 2018.