Towards Context-Aware Evaluation for Image Search

Yunqiu Shao
BNRist, DCST, Tsinghua University
Beijing, China
shaoyunqiu14@gmail.com

Jiaxin Mao
BNRist, DCST, Tsinghua University
Beijing, China
maojiaxin@gmail.com

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

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

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

ABSTRACT

Compared to general web search, image search engines present results in a significantly different way, which leads to changes in user behavior patterns, and thus creates challenges for the existing evaluation mechanisms. In this paper, we pay attention to the context factor in the image search scenario. On the basis of a mean-variance analysis, we investigate the effects of context and find that evaluation metrics align with user satisfaction better when the returned image results have high variance. Furthermore, assuming that the image results a user has examined might affect her following judgments, we propose the **C**ontext-**A**ware **G**ain (CAG), a novel evaluation metric that incorporates the contextual effects within the well-known gain-discount framework. Our experiment results show that, with a proper combination of discount functions, the proposed context-aware evaluation metric can significantly improve the performances of offline metrics for image search evaluation, considering user satisfaction as the golden standard.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**.

KEYWORDS

Image search, context, evaluation, user satisfaction

ACM Reference Format:
Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Towards Context-Aware Evaluation for Image Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331343

*Corresponding author

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).

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331343

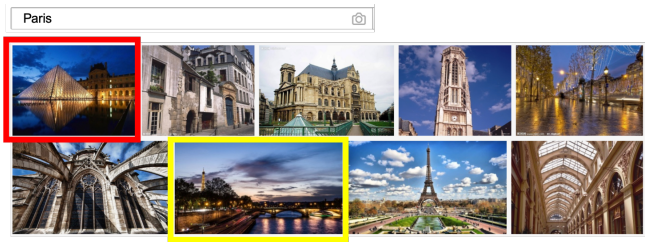


Figure 1: An example of the image result layout.

1 INTRODUCTION

With the rapid growth of multimedia contents on the web, image search has become a markedly active part within web search. Compared to general web search engines, image search engines display the search results in a different way. On the search result pages (SERPs) of image search, image results are placed in a two-dimensional panel rather than a top-down result list. Instead of snippets, snapshots of images are presented. Users can preview the image as well as the metadata without clicking in most image search engines. Figure 1 gives an example of the search result page (SERP). Due to this presentation layout, it is much easier for users to compare among image results on SERPs. The context becomes an influential factor when the user makes judgments on image items. Considering the example in Figure 1, the first image (marked by the red rectangle) is highly relevant to the query ("Paris"), and the image in the later position (marked as the yellow rectangel), although itself is also annotated as quite relevant, its relevance level seems decreased compared with the first image. In this paper, we consider the relevance of other images around one image item as its context.

Previous works attempted to improve diversity when ranking image results from the perspectives of both visual features [10] and relevance judgments [9], which also shed lights on the influences of context in the image search scenario. But how the context affects evaluation for image search is still an open question.

Evaluation sits at the center of IR research. Carterette [1] proposed a conceptual framework for model-based metrics such as Rank-Based Precision (RBP), Discounted Cumulative Gain(DCG) and Expected Reciprocal Rank (ERR). Considering the differences between general web search and image search, evaluation measures also need to be adjusted. Zhang et al. [12] compared the performances of widely-used traditional offline and online metrics in

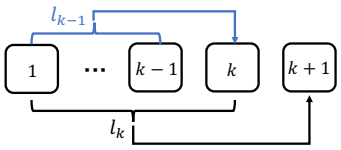


Figure 2: An illustration of context-aware user model: the relevance score of k-th result perceived by the user is affected by the results ranked in former positions, denoted by l_{k-1}.

the image search scenario and points out that the existing metrics cannot well align with user satisfaction. Previous works mainly focused on the comparisons from various angles but the context factor has not been thoroughly studied. In this paper, focusing on the evaluation of image search with the context factor considered, we investigate in the following research questions.

- **RQ1**: Does context matter for image search evaluation?
- **RQ2**: How can we consider the context factor in the evaluation of image search, and improve the performance?

In order to address these research questions, we conduct several experiments on a field study dataset [11]. We identify the context factor for image search evaluation by a mean-variance analysis. Further, we design the **C**ontext-**A**ware **G**ain (CAG), a new evaluation metric for image search, which can be easily combined with traditional evaluation metrics. Experiment results show that the proposed context-aware metric has significantly better correlations with user satisfaction in image search.

2 CONTEXT-AWARE EVALUATION

2.1 Evaluation Framework

Numbers of widely-used traditional metrics can be generalized in a gain-discount framework [1] as (1). $g(\cdot)$ characterizes the gain at position k, and it is always a function of relevance score, denoted by $g(k) = g(r_k)$, where r_k means the relevance score of the k-th result [6]. Assuming that users get less interested or are more likely to leave when scanning down the result list, $d(k)$ characterizes the discount factor of the k-th result. For example, for metric RBP with persistence parameter p, $d(k) = (1-p)\cdot p^{k-1}$ with $g(k) = g(r_k)$ set within the $[0, 1]$ range , and for metric DCG, $d(k) = 1/\log(k+1)$.

$$M = \sum_{k=1}^{K} g(k) \cdot d(k) \tag{1}$$

2.2 Context-Aware Gain

Our proposed metrics are based on the framework described above. Different from traditional web search, snapshots of image results, instead of snippets, are directly placed on the SERPs, enabling the user to compare image results more easily. Therefore, the gain that a user obtains from the k-th result is also influenced by the relevance scores of results that she has examined before, as Figure 2 shows. Using l_{k-1} to represent the result list before k-th position, we augment the gain function with l_{k-1} to incorporate the variable.

$$g(k) = g(l_{k-1}, r_k) \tag{2}$$

We further assume that a user will seek for the most relevant results, therefore, when she examines the k-th result, she would compare it with the most relevant result she has encountered. So the perceived relevance level of k-th image result is affected by the highest relevance score in l_{k-1}. We use o_{k-1} to denote the highest relevance score in the list l_{k-1}, and the relevance score of k-th result is discounted by o_k (the maximum of r_k and o_{k-1}). The adjusted relevance score of k-th result is encoded as (3), where r_k is the original relevance score of k.

$$r'_k = \frac{r_k}{o_k} \cdot r_k = \frac{r_k}{\max(r_k, o_{k-1})} \cdot r_k \tag{3}$$

Note that if $o_k = 0$, we set $r'_k = r_k = 0$. This happens when all of the images in the result list l_k are totally irrelevant, i.e. $r_j = 0$, where $j = 1, 2, ..., k$. In this study, we use 101-level relevance scores (see section 3.1), so the case is not so common.

Prior work [8] indicated that users' satisfaction may depend on a group of results rather than a single item due to the visual image panel. We use a sliding window to group the recently examined images, and use the average score to characterize the integral perception of the gain. Note that in formula (4), w denotes the window size and we only consider the first k images when $k < w$.

$$g(l_{k-1}, r_k) = \frac{\sum_{i=k-w+1}^{k} r'_i}{w} \tag{4}$$

Combing the **C**ontext-**A**ware **G**ain (CAG) with the evaluation framework, we can get our metrics in the following form.

$$M = \sum_{k=1}^{K} g(l_{k-1}, r_k) \cdot d(k) \tag{5}$$

3 EXPERIMENTAL SETUP

3.1 Dataset

Instead of traditional controlled lab experiment data, we use the field study dataset [1] collected by Wu et al. [11]. The dataset contains one-month image search logs of participants collected by a web-browser plugin. The participants were also asked to provide explicit 5-point search satisfaction feedback for each query. Fine-grained relevance annotations (ranging from 0 to 100) were gathered through crowdsourcing, and each query-image pair was annotated by at least 5 different workers, following the works of Shao et al. [8]. In summary, the original dataset contains 2,040 queries submitted by 50 participants, as well as 270,315 images with relevance scores annotated by crowdsourcing.

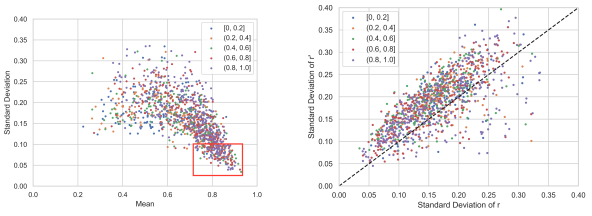
3.2 Data Cleansing

We first remove queries with over 90% invalid relevance annotations (marked as -1 or -2 in the original data). We also exclude the records of the participants who submitted fewer than 3 queries. Considering satisfaction might be quite subjective [4] and the score scales may differ with users, we normalize the satisfaction scores labeled by each user utilizing min-max scale based on formula (6), where sat_i is one satisfaction score given by one participant. $Min(Sat)$ and $Max(Sat)$ refer to the minimum and maximum value of all the

[1] http://www.thuir.cn/group/~YQLiu/

Table 1: Statistics of the dataset

	ID	# sessions	# participants	# queries	# images
≥ 10 rows	1	413	40	1,248	184,405
≥ 15 rows	2	331	39	850	152,962



(a) Mean-STD Distribution (b) STD Distributions of r and r'

Figure 3: (a) Mean and standard deviation of relevance scores of top five rows, the color of point represents normalized satisfaction scores per query. (b) Standard deviation of original relevance score r and modified relevance r' per query.

satisfaction scores the participant has labelled.

$$sat_i' = \frac{sat_i - Min(Sat)}{Max(Sat) - Min(Sat)} \tag{6}$$

As for the relevance score, we use the arithmetic average of scores given by different workers, and scale the raw relevance score to the [0, 1] range. Since browsing depths are different among queries, we only keep the queries which have no less than 10 rows of annotated images in consistent with previous works [12], which make up 66.2% of the original data. We further select queries with no less than 15 rows of annotated images to investigate the performances of metrics with deeper stopping depth. Table 1 gives the detailed statistics.

4 EXPERIMENT RESULTS

User satisfaction is widely considered as the golden standard in search evaluation [2, 5, 7, 12]. In this part, we measure how metrics align with user satisfaction. We utilize Spearman's rank correlation coefficient as our main measurement instead of Pearson's correlation coefficient, since it does not assume a normal data distribution. In this section, we first conduct a mean-variance experiment with regard to **RQ1**. As for **RQ2**, we compare the performances of our context-aware metrics with those of traditional measures according to their correlations with user satisfaction.

According to Zhang et al. [12], users tend to be quite patient and examine lots of images in the scenario of image search. Therefore, we set persistence parameter p in RBP as 0.95, which is suggested to represent patient and extremely patient users [8, 12]. Considering the number of image varies in each row, we use the number of images evaluated by metrics as the normalization factor [12].

Table 2: Spearman's rho (r_s) between user satisfaction and metrics calculated at queries of high/low variances. (* indicates the correlation is significant at $p < 0.05$ level.)

	Low-Var	High-Var
RBP (0.95)	0.127*	0.292*
DCG@10r	0.111	0.277*
CG	0.063	0.213*
AVG	0.059	0.206*
ERR	0.121*	0.288*
MAX	0.059	0.221*

4.1 Variance-Aware Evaluation

Figure 3(a) shows the mean-std distribution of image relevance scores of first five rows per query[2] along with the normalized satisfaction scores. We can observe there is a dense area in the bottom right corner, which means the images returned by the search engines are mostly highly relevant. Meanwhile, satisfaction scores of this area are mostly high as well since most of the data points in this area are purple and red. Further, we rank the queries according to the result variance in the descending order, and select the top 25% and last 25% queries respectively. We calculate the traditional evaluation metrics based on the image results of first ten rows in these two query sets. Table 2 gives the Spearman's rank correlation coefficients between metrics and user satisfaction. We find that for queries with low variance results, the evaluation metrics almost fail while metrics have better discriminative power for the high variance results. By case study, queries with low variance results usually ask for some specific items, and most of image results show the similar items, just different in angles or other decorations. Because of the display of image previews by image search engines, it is much easier for users to compare among image results. So the performances of evaluation metrics vary on different conditions of the result context. In conclusion, offline metrics calculated at high-variance results align with user satisfaction better.

4.2 Context-Aware Gain

Our context-aware evaluation mainly modifies the gain function, while leaving the discount function unchanged. In this experiment, we combine the context-aware gain with the discount functions of traditional metrics, i.e. RBP, DCG, CG, ERR, AVG, and MAX [12]. We set the window size $w = 10$ in our experiment, considering the number of images in a row varies and this window size usually contains about one or two rows. We measure the performances of evaluation metrics by comparing their correlations (r_s) with user satisfaction feedbacks, and calculate the significant level of difference between correlation coefficients with reference to Cohen [3]. In regard to **RQ2**, we conduct experiments on two datasets. We calculate metrics based on the 10 rows of image results on Dataset_1, while for Dataset_2, which contains queries along with no less than 15 rows of images, we evaluate the first 5, 10, and 15 rows of images. Table 3 gives the result.

Firstly, we observe that metrics with slower decay discount factors, like DCG and RBP align with user satisfaction better in image

[2]In most search engines, the first page contains no more than 5 rows of images

Table 3: Spearman's rho (r_s) between user satisfaction and metrics calculated based on different gain functions. ORG represents original gain and CAG represents Context-Aware Gain. (* indicates the correlation is significant at $p < 0.05$ level. † indicates the difference between r_s is significant at $p < 0.05$ level based on the same metric.)

	Dataset_1@10r		Dataset_2@5r		Dataset_2@10r		Dataset_2@15r	
	ORG	CAG	ORG	CAG	ORG	CAG	ORG	CAG
RBP (0.95)	0.281*	0.304*†	0.276*	0.308*†	0.251*	0.287*†	0.243*	0.279*†
DCG	**0.300***	**0.325***†	**0.297***	**0.323***†	**0.288***	**0.323***†	**0.282***	**0.322***†
CG	0.274*	0.303*†	0.293*	0.320*†	0.276*	0.311*†	0.265*	0.309*†
AVG	0.269*	0.303*†	0.288*	0.319*†	0.270*	0.311*†	0.258*	0.309*†
ERR	0.221*	0.221*	0.172*	0.174*	0.173*	0.175*	0.173*	0.176*
MAX	0.254*	0.262*	0.261*	0.251*	0.258*	0.251*	0.260*	0.251*

search. It indicates that the users tend to be patient to examine numbers of images when using image search engines, which is consistent with previous work [12]. Secondly, when using discounting factors of RBP, DCG, CG, and AVG, the **C**ontext-**A**ware **G**ain (CAG) always significantly outperform the original gain (ORG) on both two datasets, which verifies the benefits of context-aware gain. However, we can hardly find significant differences in ERR and MAX. For one thing, the cascade model (i.e. ERR) or the metric focused on one specific image (i.e. MAX) can not model user behavior well in image search, so the assumptions that CAG relies on fail on both metrics. For another, the context-aware gain mainly makes some corrections on the basis of the image relevance score itself. Therefore, the impacts of the context-aware gain can be accumulated and played out when using models with slower decay. Besides, we compare the STD distribution of modified relevance score r' with the that of the original relevance score r, as shown in Figure 3(b). The variance among image results have been enlarged for most of queries (large proportion of points are above the black dash line). Thirdly, the evaluation depth does not affect the performances of metrics much, which indicates that it is not very meaningful to evaluate too deep in image search. Moreover, we observe that conditioning on different evaluation depths, CAG still benefits most of evaluation metrics, and it achieves the best performances when combined with DCG in this experiment. In summary, the context-aware gain can benefit image search evaluation on the basis of traditional evaluation metrics that have a slow decay discount factor.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we mainly investigate the context factor in image search evaluation. We utilize the field study dataset that can reflect realistic user experience in image search. Focusing on the research questions, we summarize our contributions and conclusions as follows. With regard to **RQ1**, we conduct a mean-variance analysis to investigate the influences of result context and find that evaluation metrics reflect user satisfaction better when the returned image results are of high relevance variance. To address **RQ2**, we further design the context-aware gain and combine it with various discount functions. We regard user satisfaction as the gold standard and compare how metrics correlate with user satisfaction. Our experiment results show that combined with the discount function which has a slower decay or models rather patient users, e.g. DCG, RBP, our context-aware gain can achieve significant improvements in image search evaluation.

Our work is a first attempt to combine the context factor with evaluation metrics for image search. There are a few limitations that we would like to list as possible future work directions. (1) We assume the users examine image results in a sequential manner, that is to say, from left to right within a row and move to the next row after browsing an image row. Different examining patterns might be worth investigating. (2) We only combine the context-aware gain with discount factors of some existing evaluation metrics. Discount functions designed for image search are still worth for further study.

REFERENCES

[1] Ben Carterette. 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 903–912.

[2] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. (2017).

[3] Jacob Cohen and Patricia Cohen. 1983. *Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.)*. Lawrence Erlbaum Associates.

[4] Diane Kelly et al. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends® in Information Retrieval* 3, 1–2 (2009), 1–224.

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

[6] Tetsuya Sakai. 2014. Metrics, statistics, tests. In *Bridging Between Information Retrieval and Databases*. Springer, 116–163.

[7] Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. 2010. Do user preferences and evaluation measures line up?. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 555–562.

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

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

[10] Reinier H van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. 2009. Visual diversification of image search results. In *Proceedings of the 18th international conference on World wide web*. ACM, 341–350.

[11] 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 12th ACM International Conference on Web Search and Data Mining*. ACM.

[12] 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*. ACM, 615–624.