Preference-based Evaluation Metrics for Web Image Search

Xiaohui Xie
BNRist, DCST, Tsinghua University
Beijing, China
xiexh_thu@163.com

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

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

Maarten de Rijke
University of Amsterdam & Ahold Delhaize
Amsterdam, The Netherlands
derijke@uva.nl

Haitian Chen
DCST, Tsinghua University
Beijing, China
cht16@mails.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

Following the success of Cranfield-like evaluation approaches to evaluation in web search, web image search has also been evaluated with absolute judgments of (graded) relevance. However, recent research has found that collecting absolute relevance judgments may be difficult in image search scenarios due to the multi-dimensional nature of relevance for image results. Moreover, existing evaluation metrics based on absolute relevance judgments do not correlate well with search users' satisfaction perceptions in web image search.

Unlike absolute relevance judgments, preference judgments do not require that relevance grades be pre-defined, i.e., how many levels to use and what those levels mean. Instead of considering each document in isolation, preference judgments consider a pair of documents and require judges to state their relative preference. Such preference judgments are usually more reliable than absolute judgments since the presence of (at least) two items establishes a certain context. While preference judgments have been studied extensively for general web search, there exists no thorough investigation on how preference judgments and preference-based evaluation metrics can be used to evaluate web image search systems. Compared to general web search, web image search may be an even better fit for preference-based evaluation because of its grid-based presentation style. The limited need for fresh results in web image search also makes preference judgments more reusable than for general web search.

In this paper, we provide a thorough comparison of variants of preference judgments for web image search. We find that compared to strict preference judgments, weak preference judgments require less time and have better inter-assessor agreement. We also study how absolute relevance levels of two given images affect preference judgments between them. Furthermore, we propose a preference-based evaluation metric named Preference-Winning-Penalty (PWP) to evaluate and compare between two different image search systems. The proposed PWP metric outperforms existing evaluation metrics based on absolute relevance judgments in terms of agreement to system-level preferences of actual users.

CCS CONCEPTS

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

KEYWORDS

Web image search; Preference judgment; Evaluation metric; User behavior

ACM Reference Format:
Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Haitian Chen, Min Zhang, and Shaoping Ma. 2020. Preference-based Evaluation Metrics for Web Image Search. 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, 10 pages. https://doi.org/10.1145/3397271.3401146

1 INTRODUCTION

Offline evaluation in web image search relies heavily on absolute judgments of relevance to generate a ground-truth ranking of search results in response to a query [9, 25, 28]. Following the Cranfield paradigm [3], absolute judgments of relevance require assessors to determine the relevance of an image result on a graded scale, independent of any other results. However, such graded relevance judgments have a number of limitations:

(1) The lack of a universal interpretation of multi-valued relevance scales makes it hard to compare scales. For example, Yang et al. [28] label images with three levels: "irrelevant," "fair" and "relevant." Zhang et al. [30] adopt a 4-point scale relevance annotation used by a popular commercial image search engine. O'Hare et al. [14] use a simple heuristic to combine and map topical relevance and image quality to a standard 5-point PEGFB

*Corresponding author.

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

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401146

scale (Perfect, Excellent, Good, Fair, Bad). Shao et al. [19] adopt 100-point scale relevance and show that it is better than 4-point scale relevance in terms of correlation to user satisfaction.

(2) It is difficult to define clear and sufficient grades of relevance in image search scenarios. In web image search applications, factors other than topical relevance, such as attractiveness and quality of images are often included in assessing relevance [5]. Specifically, image quality depicts the artistic value of a given image, which is a vague concept and hard to define in fine-grained grades.

(3) Existing work shows that relevance-based evaluation metrics do not correlate well with user satisfaction in web image search [27] and that they do not distinguish well between different systems or ranking functions [26].

Preference judgments have been investigated as an alternative to absolute judgments in general web search [1, 8, 20]. Instead of assigning a graded relevance label to a search result, an assessor examines two documents and expresses a preference for one over the other. Hence, there is no need to explicitly determine the number of grades and define the meaning of each grade in absolute judgments. Also, by collecting preferences directly, the difficulty in distinguishing between different levels of relevance can be circumvented. Carterette et al. [1] show that it is easier for an assessor to determine a preference for one document over another than to assign a pre-defined grade to each of them. Compared to absolute judgments, preference judgments lead to better inter-assessor agreement, less time consumption per judgment and better judgment quality in terms of agreement to user clicks and satisfaction [1, 11]. Radinsky and Ailon [16] point out that these advantages come from the pairwise nature of preference judgments in which pairs of documents can mutually act as a "context," thus providing a reference for the judges.

While considerable effort has been invested in investigating preference judgements and preference-based evaluation metrics for general web search, so far there has been relatively little work on preference-based evaluation metrics for web image search. In image search, items users search for are images instead of web pages and results are typically placed in a grid-based manner rather than a sequential result list. Users can browse more results (around 15 results) than in general web search (3-4 results) at the same time without scrolling. Hence, more preference judgments may be made by user while browsing image results than while inspecting web search results. Also, the demand for result freshness is limited in web image search compared to general web search [12], which leads to better reusability of preference judgments. Similar to Radinsky and Ailon [16], Shao et al. [19] show that considering multiple images together for relevance judgments has better performance in terms of correlation to user satisfaction, meaning that judging each image in isolation is insufficient. Together, these findings motivate us to delve into preference judgments for web image search.

In this paper, we investigate preference judgments in web image search scenarios. Preference judgments can be divided into two variants on the basis of whether a "tie" option is available or not. For **strict preference judgments**, judges can only indicate whether one image is (strongly) preferred over another or vice versa. While for **weak preference judgments**, an additional "tie" option is provided, allowing judges to state that the two images are equally irrelevant or equally relevant. Figure 1 shows an example of a weak preference judgment interface (i.e., where preference judgments with ties are collected) of two image results for the query "Fortifications of Xi'an." Based on a lab-based user study, we collect

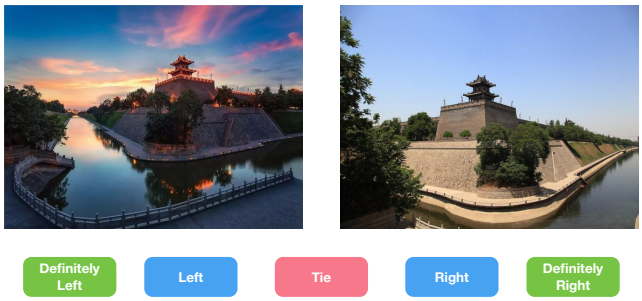

Figure 1: An example of weak preference judgment (i.e., preference judgment with tie). The given two images are results response to the query "Fortifications of Xi'an." "Definitely" here means strong preference.

preference judgment data as well as temporal information about judgments. We compare two variants of preference judgments in terms of time consumption, inter-assessor agreement, and transitivity. We find that compared to strict preference judgments, weak preference judgments lead to reduced time consumption and better inter-assessor agreement. Both types of judgment reveal high degrees of transitivity especially in judgements of results for queries with "Locate/acquire" and "Entertain" search intents. We also investigate how absolute relevance levels affect preference judgments. We find that both the relevance gap between two images and the relevance levels of two images have an effect on preference judgments. Furthermore, we propose a novel evaluation metric based on preference judgments named PWP. Extensive experiments show that the proposed PWP metric is effective and outperforms relevance-based evaluation metrics in terms of correlation with system-level preference.

In summary, we make the following contributions:

- We formally define the problem of preference judgments and preference-based evaluation for web image search. To the best of our knowledge, this paper is the first attempt to thoroughly study preference judgments in web image search.
- We conduct a lab-based user study to investigate different variants of preference judgments in web image search. Differences are observed in terms of time consumption, inter-assessor agreement, transitivity, and potential influence factors underlying judgments.
- We build and evaluate a novel preference-based evaluation metric PWP for web image search. We show that considering grid-based information and the influence of "bad" images as part of the design of evaluation metrics can be beneficial. The proposed evaluation metric PWP have better correlation with system-level preference than evaluation metrics based on absolute relevance-based judgments.
- We build a large-scale dataset with preference judgments for over 40,000 image pairs. This dataset is publicly available [1] and can be

[1] https://github.com/THUxiexiaohui/An-image-dataset-with-preference-judgments

used to further research on preference-based evaluation and to boost other types of research, e.g., to help deep neural networks to gain a better understanding of image content.

2 RELATED WORK

2.1 Absolute judgments of relevance

Similar to general web search, web image search has mostly used absolute judgments of relevance to evaluate search systems or ranking functions [14, 25, 28]. In web image search, relevance is a multi-dimensional phenomenon. Besides topical relevance, the quality of images is also used to assess relevance [5]. Existing work on web image search applies different frameworks of relevance. Different numbers of grades and different definitions of each grade are used. For example, binary relevance ("relevant" or "not relevant") is used by Sang et al. [18]. Yang et al. [28] use 3-point scale relevance judgments (i.e., "irrelevant", "fair" and "relevant") while a 4-point scale relevance used by a popular commercial image search engine is adopted in [26, 30]. Besides coarse-grained ordinal relevance judgments, a fine-grained relevance scale (S100), ranging from 0 to 100, has also been proposed and tested in web image search scenarios [19]. Compared to 4-point scale relevance judgments, Shao et al. [19] show that 100-point scale relevance judgments can help evaluation metrics to better reflect user satisfaction. O'Hare et al. [14] utilize a late-fusion approach to assess relevance. Assessors are first asked to annotate a 3-point scale topical relevance score and then a 5-point scale image quality score separately. Then, a simple heuristic that gives precedence to topical relevance is used to incorporate topical relevance and image quality and map them to a standard 5-point PEGFB scale (Perfect, Excellent, Good, Fair, Bad). On the basis of absolute relevance judgments, traditional relevance-based evaluation metrics have been used to generate a ranking score for a given result list [28, 30], e.g., normalized discounted cumulative gain (NDCG) [10] and rank-biased precision (RBP) [13]. Furthermore, Xie et al. [27] propose grid-based evaluation metrics by considering grid information to revise traditional list-based metrics.

Given that relevance judgments in web image search may have substantially different settings, it is not clear how compatible or comparable they are. Also, existing work shows that in some cases relevance-based evaluation metrics do not reflect user satisfaction and distinguish between different systems in image search scenarios [27, 30].

Our contributions in this paper complement existing work on absolute judgments of relevance for web image search by adding an investigation into preference judgments, which are an alternative to (graded) relevance judgments.

2.2 Preference judgments

Instead of considering each result in isolation and assigning a predefined grade, preference judgments consider a pair of results and ask assessors to state their relative preference. Previous work on preference judgments in information retrieval mainly focuses on general web search. Rorvig [17] shows that substituting the usual relevance judgments with preference judgments is beneficial for tasks whose goal is to find highly-relevant documents. Yao [29] study the concept of user preference based on decision and measurement theories.

There are two variants of preference judgments, i.e., *strict* preference judgments and *weak* preference judgments. Carterette et al. [1] perform the first investigation into a comparison of strict preference judgments to absolute judgments. Compared to absolute judgments, strict preference judgments lead to better inter-assessor agreement and less time consumption per judgment. Song et al. [20] and Hui and Berberich [7] investigate weak preference judgments (i.e., preference judgments with an additional "tie" option) and show effective methods to reduce the number of judgments required. Recently, Hui and Berberich [8] have compare two variants of preference judgments in terms of transitivity, time consumption and judgment quality.

Although the aforementioned work shows that preference judgments are a good alternative to (graded) absolute judgments, this work is conducted on general web search scenarios rather than for web image search. Whether the findings obtained so far also hold for web image search, which differs a lot from general web search [24], is where we contribute. In particular, we investigate preference judgments for image results and show properties (e.g., time consumption and inter-assessor agreement) of different variants of preference judgments in image search. We also examine factors that affect preference judgments.

2.3 Preference-based evaluation metrics

Using preference judgments poses a new challenge to evaluation measures since frequently used measures such as, e.g., NDCG, which are based on absolute relevance judgments, are not applicable. Carterette et al. [1] propose two preference-based evaluation metrics for general web search, i.e., precision of preferences (Ppref) and weighted precision of preferences (Wpref). Ppref measures the proportion of pairs that are correctly ordered by the search engine while Wpref is a weighted version of Ppref, which adopts a rank-based weighting scheme. It has been shown that Wpref correlates well with NDCG [1]. Yao [29] defines "distance" between a given ranking and a ground-truth ranking to form a new measure of system performance named distance-based Performance Measure (Dpm). Dpm is based on preference judgments and uses the relative order of documents. Furthermore, Chandar and Carterette [2] propose a model-based measure using preferences to assess the effectiveness of systems for the novelty and diversity task.

However, these evaluation metrics are developed for general web search. Since web image search differs from general web search in terms of result placement, search intent and interaction mechanism [23], these evaluation metrics may not be appropriate. Xie et al. [27] show that considering the grid-based result presentation-based mode used in image search as part of the design of evaluation metrics is essential.

What we add on top of prior work on preference-based evaluation metrics is that we incorporate grid-based assumptions and inter-system preference judgments to propose novel preference-based evaluation metrics for web image search.

3 UNDERSTANDING PREFERENCE JUDGMENTS

In order to get a better understanding of preference judgments in web image search, we conduct a laboratory user study to collect preference judgment data. Based on the data we collect we investigate variants of preference judgments (i.e., strict and weak) in web image search. We seek to answer the following research questions:

(RQ1) What is the difference between weak preference judgments and strict preference judgments in terms of time consumption and assessor agreement?

(RQ2) Do weak and strict preference judgments for image results exhibit transitivity, so that, e.g., image I_1 is preferred over image I_3 whenever I_1 is preferred over I_2 and I_2 over I_3?

(RQ3) What is the relationship between relevance levels of two images and the preference judgments between them?

3.1 Data collection procedure

We sample queries and images from an image search dataset consisting of data collected from a one-month field study, which is publicly available [22]. We use this dataset for two main reasons: (1) In this dataset, queries are issued by real search users and image results are returned by a commercial image search engine. This dataset has been used by previous work on evaluation metrics for web image search [27]. (2) Abundant information, including users' search intents and query-image relevance scores, is available, which enables research into factors that may affect users' preference judgments. Specifically, we sample 12 queries that cover different search intent categories according to the intent taxonomy of web image search proposed by Xie et al. [23]. The number of queries is the same as used in a previous preference study for general web search [8]. Detailed descriptions of the query set are shown in Table 1. In the field study dataset released by Wu et al.

Table 1: Description of the query set used for preference judgments. Queries are sampled from a field study dataset [22].

Search intent	Query ID	Query
Locate/Acquire	0–3	Snicker sticker/ HD Iqiyi Logo/ Diamond wallpaper/ Fudan University logo
Explore/Learn	4–7	Where is the dragon fruit growing/ Thailand map/ Jump shot training/ Purple highlight to hair
Entertain	8–11	Produce 101/ Putin riding a horse/ 2018 world cup mascot/ Louis Koo at Hong Kong Film Award

[22], each image result is assigned a 100-point scale relevance score by assessors. We divide the relevance scale into three intervals (i.e., $[0, 100/3]$, $[100/3, 100*2/3]$, $[100*2/3, 100]$) and randomly sample 4 images from each interval for each query. In this manner we obtain 12 queries, 144 image results and 792 image pairs for preference judgments.

In order to explore and compare variants of preference judgments in web image search, we recruit ten assessors to provide preference judgments. Five are asked to perform weak preference judgments (Definitely left, Left, Tie, Right, Definitely right), while the remaining assessors are asked to perform strict preference judgments (Definitely left, Left, Right, Definitely right). For each judgment, a query and two image results for this query are displayed. We randomly place these two images on two sides (i.e., left and right) to avoid position bias towards a particular side. Besides preference judgment data, the time spent on each judgment is also recorded. To make sure the time is only affected by intrinsic properties of preference judgments instead of difficulties of understanding the meaning of queries caused by differences in prior knowledge of the assessors, we show our assessors the queries as well as corresponding descriptions before the annotation procedure.

3.2 Comparison between different variants of preference judgments

We first compare strict preference judgments and weak preference judgments in terms of time-consumption. Recall that for each image-image judgment, we have five assessors to provide preference annotations. We define the time-consumption of a preference judgment to be the median value of time spent by these five assessors. Here, we use the median to aggregate times instead of the mean to reduce effects of outliers as in [24]. Then, for each query, we calculate the average time needed for preference judgment by averaging across all image-image pairs of this query. The results for the 12 queries we sampled are shown in Figure 2. From Figure 2,

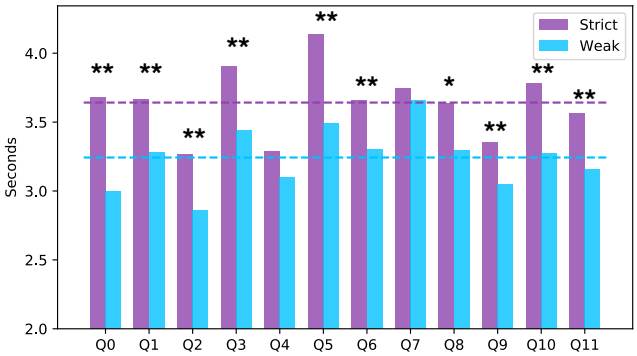

Figure 2: Average time (seconds) needed to produce strict preference judgments and weak preference judgments for different queries. ** (*): The difference is significant with p-value < 0.01 (0.05). Dashed lines refer to the average time (averaged over all image pairs).

we can see that, compared to strict preference judgments, weak preference judgments are less time consuming for each query. By performing a paired two-tailed t-test, we show that the difference between strict preference and weak preference is significant for most queries except $Q4$ and $Q7$. In Figure 2 we also display the average time required to produce the two types of preference judgment (averaged over all image pairs), using the two dashed lines. Producing weak preference judgments requires 11% less time than producing strict preference judgements ($(3.64 - 3.24)/3.64 \approx 11\%$).

Next, we calculate Fleiss' kappa scores [4] to reveal assessor agreements for different variants of preference judgments. Besides

showing results of the original preference judgments, we also show kappa scores of judgment results obtained by grouping the original ones. Specifically, we group "Definitely left" and "Left" into a single group ("preference for the left image") and "Definitely right" and "Right" too ("preference for the right image"). In this manner, strict preference with four options ("Strict (4)") can be mapped to strict preference with two options ("Strict (2)"); and weak preference with five options ("Weak (5)") can be mapped to weak preference with three options ("Weak (3)"). The Fleiss' kappa scores for each of these options are shown in Table 2. From Table 2, we see that the

Table 2: Assessor agreement of preference judgments under different search intents. The numbers in brackets are the number of options for preference judgments. By considering "Definitely left/right" and "left/right" together as preferring the left/right image, we map the four options available for strict preferences (Strict (4)) to two options (Strict (2)), and the five options available for weak preferences (Weak (5)) to three options (Weak (3)).

Fleiss' kappa	Strict (4)	Weak (5)	Strict (2)	Weak (3)
Locate/Acquire	0.308	0.302	0.386	0.554
Explore/Learn	0.241	0.286	0.403	0.487
Entertain	0.372	0.364	0.431	0.576
All	0.326	0.322	0.419	0.539

original Strict (4) preference judgments and Weak (5) preference judgments reveal similar assessor agreement (fair agreement). The Fleiss' kappa scores of the two collapsed variants of preference judgments both achieve improvements. Compared to the collapsed Strict (2) preference judgments, the Weak (3) judgments have more options but they show better assessor agreement, with a Fleiss' kappa score of 0.539 (substantial agreement). Furthermore, we see that compared to "Locate/Acquire" and "Entertain," the "Explore/learn" category has lower assessor agreement for all preference judgments. The reason can be that queries and their corresponding results under the "Explore/learn" intent are relatively difficult to understand, which might confuse assessors in some cases.

Answer to RQ1. Compared to strict preference judgments, the creation of weak preference judgments is less time consuming (about 11%) and they have better a assessor agreement (substantial agreement while considering judgments with three options (Weak (3)) in image search. In general web search, weak preference judgments also require less time-consumption compared to strict preference judgments. However, the inter-assessor agreement of strict preference judgments is higher than of weak preference judgments in general web search which is different from aforementioned findings in web image search [8], which may be caused by the intrinsic differences between web pages and image results.

3.3 Transitivity of preference judgments

Next, we examine transitivity of both strict and weak preference judgments. Transitivity of preference judgments in image search matters for multiple reasons: (1) Relevance is a multi-dimensional phenomenon in image search. Thus, assessors might provide preference judgments based on different aspects of relevance, which may violate transitivity. (2) Transitivity is a prerequisite for sorting algorithms that can be applied to reduce the number of preference judgments (from $O(N^2)$ to $O(N \log N)$) [1, 20].

Table 3: Transitivity over aggregated judgments of different assessors. We report the ratio of transitive triples out of triples in different types.

Transitivity	Strict	Weak			
	asym	asym	s2a	s2s	All
Locate/Acquire	97%	98%	96%	75%	94%
Explore/Learn	93%	93%	83%	71%	86%
Entertain	95%	97%	91%	82%	94%
All	95%	96%	91%	79%	91%

We investigate transitivity on the basis of aggregated judgments as in [8]. Majority voting is used to aggregate judgments of all assessors to generate a preference tag for a given image-image pair. We use the Strict (2) and Weak (3) preference judgments described in Section 3.2. We write $I_1 > I_2$ to denote that image I_1 is preferred over image I_2, and $I_1 \approx I_2$ if there is a tie between them. Then, a triple $\langle I_1, I_2, I_3 \rangle$, where I_1, I_2 and I_3 are images returned for a query in our dataset, is said to exhibit transitivity if it follows the definition of transitivity introduced in Section 3.1. For example, if either $I_1 > I_2, I_2 > I_3, I_1 > I_3$ or $I_1 \approx I_2, I_2 \approx I_3, I_1 \approx I_3$ hold, the triple $\langle I_1, I_2, I_3 \rangle$ exhibits transitivity.

According to [6], transitivity can be decomposed based on how many "tie" judgments exist in a triple. In particular, the "Left" and "Right" options are referred to as asymmetric relationships and the "Tie" option is referred to as a symmetric relationship. Transitivity can then be categorized as: "asym": no "Tie" judgment in a triple; "s2a": only one "Tie" judgment in a triple; "s2s": at least two "Tie" judgments in a triple. We examine how many image triples exhibit transitivity and show the ratio of transitive triples out of triples in different types in Table 3. Of course, strict preferences can only have the "asym" transitivity since they do not have the "Tie" option. From Table 3, we have following findings:

(1) Transitivity holds for more than 90% of the triples for the two variants of preference judgments. For strict preference judgments, transitivity holds for 95% of the triples, which is close to the number 96% reported in web search scenarios [8]. For weak preference judgments, the number of triples for which transitivity holds is 91% which is slightly lower than for strict preference judgments.
(2) When examining transitivity of weak preference judgments of different types (i.e., "asym", "s2a" and "s2s"), we see that "asym" transitivity holds more frequently than for strict preferences. Although "s2s" transitivity in image search holds more frequently than in web search, where it is less than 60% [8], it is less frequent than "asym" and "s2a" transitivity in image search. Hence, given three images I_0, I_1, I_2, one should be careful to conclude that $I_1 \approx I_2$ when $I_0 \approx I_1$ and $I_0 \approx I_2$ are observed.
(3) Compared to the "Locate/Acquire" and "Entertain" intents, the "Explore/Learn" intent shows weaker transitivity, which may be caused by the difficulties of "Explore/Learn" tasks as discussed in Section 3.2.

Answer to RQ2. Both strict preference judgments and weak preference judgments reveal substantial degrees of transitivity, although one should be cautious to use transitivity when two "Tie" judgments are observed in a triple. The aforementioned findings are also observed in general web search [1, 8].

3.4 Relationship between absolute judgments and preference judgments

We first define two variables examined in this section:

- **Judgment time**: Time for judging an image-image pair. We use the median value of assessors' judgment time to generate aggregated judgment times of a given image-image pair as in Section 3.2.
- **Judgment strength**: Strength of the preference towards the image with higher relevance value in a given image-image pair. We assume that we put the image with higher relevance value at the right side. We first map each assessor's preference judgment to a numeric scale ("Definitely Left" and "Left": 0; "Tie": 1; "Right" and "Definitely Right": 2) and average the scale number across assessors to obtain the judgment strength of the given image pair.

Given an image pair, we assume that image A (B) is always the image with the lower (higher) absolute relevance score and evenly divide the 100-point relevance scale into 10 intervals from interval 1 ([0,10]) to interval 10 ([90,100]). Five types of relevance relationship between two images are examined in this paper:

(1) Relevance gap of two images: Rel(B) − Rel(A);
(2) Ratio between the relevance value of two images: Rel(B)/Rel(A);
(3) Overall relevance value of the pair of images: Rel(B) + Rel(A);
(4) The higher relevance value in the pair of images: Rel(B);
(5) The lower relevance value in the pair of images: Rel(A).

In Figure 3, we plot the judgment time distribution and the judgment strength distribution of different variants of preference judgment. We show correlations between different types of relevance relationship between two images and judgment time (cost) and judgment strength in Table 4 and Table 5, respectively. In this paper, we use both a parametric test (Pearson r correlation) and a non-parametric test (Spearman correlation) to show the level and significance of correlations.

Table 4: Correlation between relevance levels of two images and the judgment time (cost) of preference judgments between them. ** (*): The correlation is significant with p-value < 0.01 (0.05).

Time (Cost)	Pearson r		Spearman r	
	Strict	Weak	Strict	Weak
Rel(B) − Rel(A)	-0.235^{**}	-0.061	-0.232^{**}	-0.041
Rel(B)/Rel(A)	-0.160^{**}	-0.051	-0.218^{**}	-0.100^{**}
Rel(B) + Rel(A)	0.003	0.172^{**}	0.017	0.169^{**}
Rel(A)	0.126^{**}	0.181^{**}	0.127^{**}	0.160^{**}
Rel(B)	-0.118^{**}	0.114^{**}	-0.083^{*}	0.138^{**}

Based on Figure 3, Table 4 and 5, we have following observations:

(1) The difference (gap or ratio) between the relevance value of two images correlates better with judgment time of strict preference than of weak preference. We can see from Figure 3(a) that when the difference becomes smaller (closer to the diagonal of the heat map), the time to produce strict preference judgments becomes longer.
(2) The overall relevance value of two images (Rel(B)+Rel(A)) or the relevance value of each image (Rel(B)/Rel(A)) correlates better with judgment time of weak preference than of strict preference. Although the correlation is significant, it is relatively weaker than results of difference (gap or ratio).
(3) Both judgment strength of strict preference and weak preference have a substantial correlation with the relevance gap while the correlations with the ratio are moderate. Figure 3(c) and Figure 3(d) indicate that when the difference between relevance values of two images becomes larger (closer to the origin of coordinates), the judgment strength of the preference becomes larger.
(4) The judgment strength of weak preference has a stronger correlation with the overall relevance value of two images than strict preference. An interesting finding is that the relevance value of the image with the higher value among the image pair (i.e., Rel(B)) has a strong correlation with the judgment strength of preference judgments, which might indicate that images with high relevance value have a strong effect on users' preference.

These findings share the intrinsic consistency with the psychophysical regularity revealed in [15, 21]. Weber [21] discovered that the probability that a subject will make the right choice as to which of two slightly different weights is heavier only depends on the ratio between the weights, which is now known as Weber's Law. Pardo-Vazquez et al. [15] indicated that decision times to discriminate between two sounds of slightly different intensities and the loudness of the pair of sounds are linked.

Table 5: Correlation between relevance levels of two images and the judgment strength of preference judgments between them. ** (*): The correlation is significant with p-value < 0.01 (0.05).

Strength	Pearson r		Spearman r	
	Strict	Weak	Strict	Weak
Rel(B) − Rel(A)	0.495^{**}	0.551^{**}	0.546^{**}	0.587^{**}
Rel(B)/Rel(A)	0.244^{**}	0.267^{**}	0.428^{**}	0.388^{**}
Rel(B) + Rel(A)	0.178^{**}	0.267^{**}	0.017	0.331^{**}
Rel(A)	-0.105^{**}	-0.056	-0.118^{**}	-0.002
Rel(B)	0.404^{**}	0.508^{**}	0.349^{*}	0.532^{**}

Answer to RQ3. Analogous to findings in psychophysical experiments [15, 21], we find that the difference between the relevance value of two images correlates well with judgment time of strict preference and judgment strength of both strict preference and weak preference. Furthermore, results show that the overall relevance value of two images and the relevance value of the image that is the more relevant one in an image pair, correlate well with the judgment strength of weak preferences.

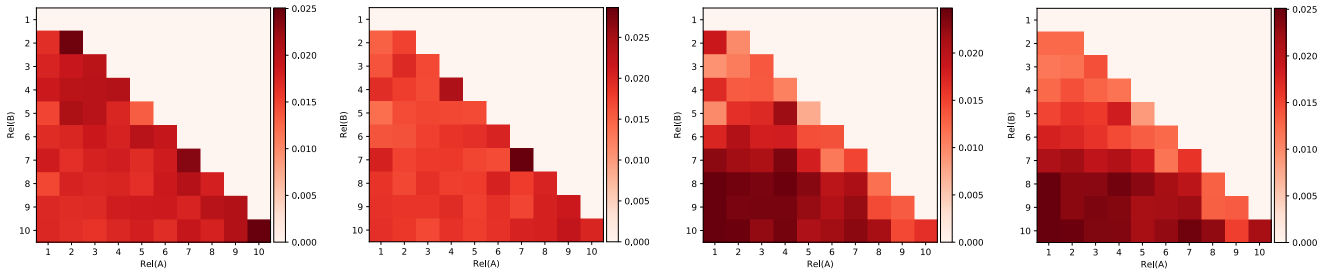

Figure 3: Judgment time distribution (seconds) of strict (a) and weak (b) preference judgments. Judgment strength distribution of strict (c) and weak (d) preference judgments. Rel(A)/Rel(B) refers to the relevance score of the image with the lower/higher relevance level among an image pair (A,B).

4 PREFERENCE-BASED EVALUATION METRICS

In this paper, we aim to use preference judgments to evaluate web image search engines by proposing a novel preference-based evaluation metric. Specifically, given a pair of result sets returned by two different search systems, denoted by I_{S_1} and I_{S_2} respectively, we propose a metric PWP that takes preference judgments for image pairs extracted from $I_{S_1} \cup I_{S_2}$ as input and outputs a probability that system S_1 is preferred to system S_2. We define $\omega(I_S)$ as a set of image pairs in which images are extracted from the result set I_S under certain assumptions. Then the problem investigated in this paper can be formally defined as:

$$P(S_1 \triangleright S_2) = f(\omega(I_{S_1} \cup I_{S_2})). \tag{1}$$

where $\triangleright$ denotes the pairwise preference between two items (e.g., systems or image results); $\triangleright$ can be represented as $\succ$ for strict preference and as $\succcurlyeq$ for weak preference. We construct f based on three parts, i.e., Preference Matching Rate (*PMR*), Winning rate (*WR*) and Penalty for bad cases (*PB*). While *PMR* measures preference inside a system, *WR* and *PB* take preference judgments for images from different systems into consideration. These are crucial for the task discussed in our paper.

4.1 Preference Matching Rate (*PMR*)

PMR is the proportion of pairs that are correctly ordered by the system. To obtain this measure, two aspects should be considered, i.e., what is the right order and how many results? The first aspect focuses on placing results one by one following a certain order according to users' preferences while the second aspect determines the scale of results used to extract image pairs. In web image search, grid-based result panels are used instead of the linear result lists that are common in general web search. Hence, rather than a linear order that places preferred results from top to bottom, users' examination sequence on grid-based search engine result pages (SERPs) of image search should be defined. With a pre-defined examination sequence, we can map the position (row, column) of a given image result i to a numerical value denoted by d_i as in [25]. For example, consider a SERP consisting of 3 rows of results, where each row has 4 results. Assuming a default examination sequence that is from left to right and top to bottom, then the rank of the second result in the second row is 6 (mapped from (2, 2)).

With this definition of rank in grid-based panels, the preference matching rate of a given system S can be written as:

$$PMR(S) = \frac{\sum_{\langle i,j\rangle \in \omega(I_S)} (\mathbb{1}_{d_i < d_j} \mathbb{1}_{i \triangleright j} + \mathbb{1}_{d_i > d_j} \mathbb{1}_{i \triangleleft j})}{|\omega(I_S)|}, \tag{2}$$

where $|\cdot|$ denotes the number of items in the set and $\mathbb{1}$ is the indicator function. The numerator represents the number of correctly ordered pairs in terms of users' preference judgments. *PMR* depicts the performance of a search system in terms of placing preferred results at positions receiving more attention.

Xie et al. [27] show that considering grid information as part of the design of evaluation metrics can be beneficial. We also incorporate two grid-based assumptions, i.e., "middle position bias" and the "nearby principle" introduced in [24], into the preference matching rate. Specifically, we test four variants of *PMR* in this paper:

- **Default (PMR_D)**: Users follow a left to right and top to bottom sequence. PMR_D is analogous to Ppref introduced in [1].
- **Weighted (PMR_W)**: PMR_W is a weighted version of PMR_D. For image results at ranks i and j such that $j > i$, we set the weight $w_{ij} = \frac{1}{\log_2(j+1)}$. Then PMR_W is the sum of weights w_{ij} over pairs i, j such that i is preferred to j and the rank of i is less than the rank of j. The sum of all weights w_{ij} is used as the normalizing constant. PMR_W is analogous to Wpref introduced in [1].
- **Middle position bias (PMR_M)**: In the vertical direction, users follow a top-down pattern. In the horizontal direction, users pay more attention to the middle position of a certain row. In that regard, for a given image pair $\langle i, j\rangle$ in one row, $d_i < d_j$ denotes that the image i is closer to the middle position than the image j.
- **Nearby principle (PMR_N)**: PMR_N follows the same examination sequence assumption as PMR_D. We incorporate the "Nearby principle" assumption by assuming that users will only be concerned with the relationship between two image results that are close to each other. We define a distance function D to depict the distance between two images as in [25]:

$$D = \max(|r_i - r_j|, |c_i - c_j|), \tag{3}$$

where r refers to the row number and c refers to the column number of image results. Two images will be paired in $\omega(I_S)$ when $D \leq 2$.

4.2 Winning rate (WR)

Given image sets I_S and $I_{S'}$ from two different systems S and S', respectively, the winning rate of system S can be represented as:

$$WR(S \mid S') = \frac{\sum_{i \in I_S, j \in I_{S'}} \mathbb{1}_{i \triangleright j}}{|I_S| \cdot |I_{S'}|}. \quad (4)$$

WR depicts the performance of a search system in terms of retrieving more images with high quality compared against another search engine.

4.3 Penalty for bad cases (PB)

Finally, we consider the effect of bad cases on users' preference judgments for two search systems. Since the proposed evaluation metric is purely preference-based, instead of asking assessors to provide bad case judgments as in [1], we define "bad case" in the following preference-based manner: Given two search systems S and S', an image result i in I_S is a bad case if, and only if, $j \triangleright i$ for any image result j in $I_{S'}$. Furthermore, we define a penalty parameter γ and represent the penalty for bad cases as:

$$PB(S \mid S') = \gamma^n, \quad (5)$$

where n is the number of bad cases in system S. PB is used to refine preference towards a search system retrieving "bad case".

4.4 Preference-Winning-Penalty (PWP)

We then combine these three parts — PMR, WR, and PB — using a simple heuristic: when comparing two search systems S_1 and S_2, users will consider preference judgments for image pairs both inside S_1 and between S_1 and S_2 to generate a preference score for S_1. However, the appearance of bad cases in S_1 may discount the preference towards S_1. Given the competitive system S_2, users' preference towards the system S_1 can be represented as:

$$PWP(S_1 \mid S_2) = (\lambda PMR(S_1) + (1-\lambda) WR(S_1 \mid S_2)) \cdot PB(S_1 \mid S_2), \quad (6)$$

where λ is a trade-off parameter to combine the preference matching rate and winning rate of system S_1 given system S_2. Hence, to obtain the preference score for a certain system, there are two hyper-parameters that need to be determined, i.e., the trade-off parameter λ and the penalty parameter γ. We will discuss the effect of different value of these two parameters in Section 5. Then, the preference-based evaluation metric $\overline{PWP}$ can be represented as:

$$\overline{PWP} := f_{PWP} = \frac{1}{1 + e^{PWP(S_1|S_2) - PWP(S_2|S_1)}}. \quad (7)$$

We use a sigmoid function to combine two preference scores together and leave other, more sophisticated ways to combine these two scores as future work. Moreover, we can test the performance of different parts of PWP by replacing PWP in Eq. 7 with corresponding parts, e.g., $\overline{PB}$ (i.e., f_{PB}).

To avoid ambiguity, we use $\overline{M}$ to denote a preference-based metric that uses M to compute preference score towards a given system.

5 EVALUATING PREFERENCE-BASED EVALUATION METRICS

We first introduce the experimental setup including the dataset and baseline models used in our experiments aimed at investigating the effectiveness of the proposed evaluation metrics and how different parameters (λ and γ) affect their performance. We then show experimental results and provide a detailed analysis. Both the dataset and the code for our evaluation metrics will be made publicly available upon publication of the paper.

Table 6: Statistics of the dataset used in our experiments. The three numbers in brackets refer to the number of times of Sogou wins, the two systems tie, and Baidu wins, respectively, in terms of SERP-level preference.

Dataset	#Queries	#Images	#Image pairs
Search engines	102 (29, 46, 27)	2,919	41,538

5.1 Experimental setup

5.1.1 Dataset. We form our dataset using the following procedure: (1) We randomly sample 150 torso queries from a search log in October 2017 from the Sogou image search engine,[2] which is popular in China. (2) We use these queries to obtain the top three rows of image results (results users can view before scrolling under most resolution settings of the web browser) returned by two popular search engines in China, i.e., Sogou and Baidu.[3] (3) We discard pornographic queries and queries of which advertising results are shown on SERPs, which leaves us with 102 queries and 2,919 images. (4) We recruit assessors to provide preference judgments for image pairs both from one of the two search engines and between two search engines. For each image pair out of 41,538 image-image pairs, three assessors are recruited. Since compared to strict preference judgments, weak preference judgments require less time and higher inter-assessor agreement, we use weak preference judgments for this large-scale annotation effort. The Fleiss' kappa for weak preference judgments with three options is 0.605 (substantial agreement). Majority voting is used to aggregate preference judgments of different assessors. (5) Besides preference judgments, we also recruit three assessors (different from the assessors performing preference judgments) to provide 100-point relevance annotations for 2,919 images. Shao et al. [19] shows that fine-grained relevance judgments (100-point) are better than coarse-grained relevance judgements (4-point) in terms of reflecting user satisfaction. (6) We ask 5 professional assessors to provide weak preference judgments for 102 SERP pairs, which we call *SERP-level preferences* (with possible outcomes "Sogou wins," the two systems tie, or "Baidu wins"). SERP-level preferences are used as the golden standard to evaluate the proposed evaluation metrics.

The statistics of the dataset used in our experiments are shown in Table 6.

5.1.2 Evaluation metrics. We compare the proposed evaluation metrics against relevance-based evaluation metrics as well as existing preference-based evaluation metrics in terms of the correlation with SERP-level preference.

For relevance-based evaluation metrics, we evaluate the performance of two widely-used metrics, NDCG and RBP. We show results of NDCG with different cut-offs (10 and 15) and RBP with

[2] http://sogou.com

[3] http://baidu.com. Sogou and Baidu are among the top 3 popular search engines in China in recent years.

different persistency parameters ($\rho = 0.99, 0.8$) as in [19]. As formulated by Eq. 7, we calculate metric scores of two systems, respectively, and then use a sigmoid function to combine these two scores together.

For preference-based evaluation metrics, the proposed $\overline{PWP}$ and other variants of f, e.g., $\overline{PMR}$ are evaluated. Recall that $\overline{PMR_D}$ and $\overline{PMR_W}$ are grid-based versions of existing preference-based evaluation metrics, Ppref and Wpref, that have previously been developed for general web search. Since we use weak preference judgments to build our metrics, we should specify the preference relationship $\triangleright$ used in Section 4. Specifically, the preference relationship is $\succcurlyeq$ in Eq. 2 as in [1] and is $\succ$ in Eq. 4. We conduct comprehensive experiments to show the performance of the aforementioned preference-based evaluation metrics as well as how different hyperparameters (i.e., λ and γ) affect their performance.

5.2 Results and analysis

In the first experiment, we evaluate the performance of different variants of the preference matching rate (i.e., $\overline{PMR_D}$, $\overline{PMR_W}$, $\overline{PMR_M}$, $\overline{PMR_N}$) to show whether incorporating grid information is beneficial. Results are shown in Table 7. From Table 7, we can observe that

Table 7: Correlation between different variants of the preference matching rate ($\overline{PMR_D}$, $\overline{PMR_W}$, $\overline{PMR_M}$, $\overline{PMR_N}$) and SERP-level preferences.

	$\overline{PMR_D}$	$\overline{PMR_W}$	$\overline{PMR_M}$	$\overline{PMR_N}$
Pearson r	0.255	0.250	0.244	0.260
Spearman r	0.226	0.225	0.210	0.243

the difference between Ppref ($\overline{PMR_D}$) and Wpref ($\overline{PMR_W}$) in image search scenarios is not significant. As to the two grid-based assumptions, while "Middle position bias" ($\overline{PMR_M}$) shows no benefit when comparing between two systems, the "Nearby principle" ($\overline{PMR_N}$) can improve the performance of the default version of $\overline{PMR}$ ($\overline{PMR_D}$), which indicates that considering preference judgments following the "Nearby principle" as part of the design of preference-based evaluation metrics is beneficial.

In the second experiment, we delve into the effect of the parameter λ to see whether combining intra-system and inter-system preference judgments can further improve the performance of preference-based evaluation metrics. We use PW to denote $\lambda PMR + (1-\lambda)WR$. We show the correlation between metric scores of $\overline{PW}$ and SERP-level preferences in Figure 4. We only show results in terms of Pearson r since observations with Spearman r are similar. From Figure 4, we can see that for all variants of PMR, $\overline{PW}$ with λ with non-zero values outperform WR and PMR (0.260 reported in Table 7) which indicates that considering PMR and WR together is better than considering one facet only. Also, we can see that for all values of λ, $\overline{PW}$ with $PW = \lambda PMR_N + (1-\lambda)WR$ performs the best compared to other combinations. The best performance in Figure 4 is observed when $\lambda = 0.7$ using PMR_N. In the following experiment, we apply the best settings of the combination of PMR and WR, i.e., $0.7PMR_N + 0.3WR$.

In the third experiment, we evaluate the performance of the proposed evaluation metric $\overline{PWP}$. We aim to investigate whether

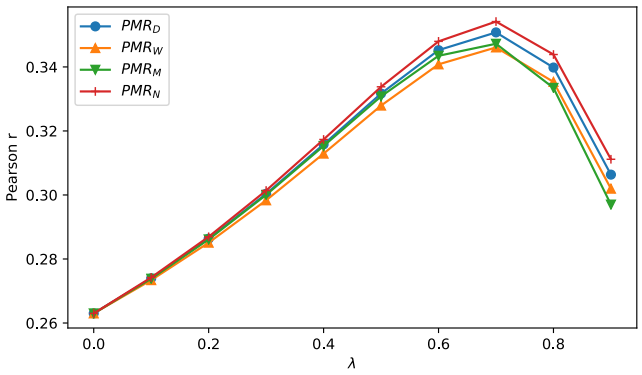

Figure 4: Correlation between metric scores of $\overline{PW}$ and SERP-level preferences under different variants of PMR (i.e., PMR_D, PMR_W, PMR_M and PMR_N) and different value of λ. When $\lambda = 0$, PW degrades to WR. ($PW := \lambda PMR + (1-\lambda)WR$)

preference-based evaluation metrics outperform relevance-based evaluation metrics in the task of comparing two search systems. We show the correlation between different evaluation metrics ($\overline{PWP}$, $\overline{PB}$, NDCG, RBP) and SERP-level preferences in Figure 5.

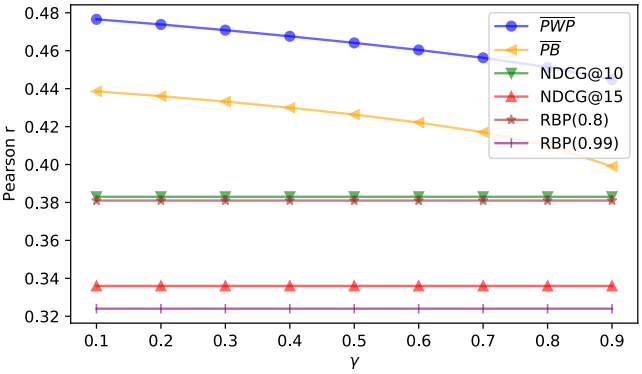

Figure 5: Correlation between different evaluation metrics and SERP-level preferences. Here, $PWP = (0.7 \cdot PMR_N + 0.3 \cdot WR) \cdot \gamma^n$ and $PB = \gamma^n$.

From Figure 5 we have following main findings:

(1) The best correlation is obtained by the preference-based evaluation metric $\overline{PWP}$ with $PWP = (0.7 \cdot PMR_N + 0.3 \cdot WR) \cdot 0.1^n$. The improvement is over 23% compared to the best performance of relevance-based evaluation metrics obtained by NDCG@15. Hence, it is promising to use preference-based evaluation metrics to evaluate image search engines.

(2) $\overline{PB}$ also shows promising results, but for all values of the penalty parameter γ, $\overline{PWP}$ is better than $\overline{PB}$ in terms of Pearson r, which indicates that PMR and WR are also crucial and combining three parts together can achieve better results.

(3) A small value of γ is better than a large value of γ for both $\overline{PWP}$ and $\overline{PB}$. Recalling that γ is the penalty parameter of which a smaller value refers to a heavier penalty. Hence, we can conclude that the appearance of "bad cases" has a significant effect on users' preference judgments between two search systems.

6 CONCLUSION

In this paper, we have studied preference judgments in web image search scenarios. We have investigated two variants of preference judgments, i.e., strict preference judgments and weak preference judgments and obtain the following insightful findings: (1) Compared to strict preference judgments, the creation of weak preference judgments is less time consuming (around 11%) and have a better inter-assessor agreement. (2) Both strict preference judgments and weak preference judgments exhibit substantial degrees of transitivity (over 90% transitive triples). (3) The difference (gap or ratio) between the absolute relevance values of two images correlates well with judgment time of strict preference and judgment strength of both strict and weak preference.

We have also proposed a preference-based evaluation metric named *PWP* that combines preference matching rate, winning rate and penalty for bad cases. Our experimental results show that the proposed preference-based metric outperforms existing relevance-based metrics, e.g., NDCG and RBP in terms of correlation to SERP-level preferences.

Limitations of the proposed preference-based evaluation metric which may guide future work include the following: (1) Preference-based evaluation require a larger number of judgments than relevance-based evaluation even after assuming transitivity. How to reduce the number of judgments without affecting the effectiveness of the preference-based metric? (2) We have only examined the performance of the propose preference-based evaluation metric in terms of correlation to SERP-level preferences. More experiments, e.g., correlation to user click and satisfaction in online settings, are left as future work.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011, 61902209), Beijing Academy of Artificial Intelligence (BAAI) and the Innovation Center for Artificial Intelligence (ICAI). All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Ben Carterette, Paul N Bennett, David Maxwell Chickering, and Susan T Dumais. 2008. Here or There. In *European Conference on Information Retrieval*. Springer, 16–27.

[2] Praveen Chandar and Ben Carterette. 2013. Preference based Evaluation Measures for Novelty and Diversity. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 413–422.

[3] Cyril Cleverdon. 1967. The Cranfield Tests on Index Language Devices. In *Aslib proceedings*, Vol. 19. MCB UP Ltd, 173–194.

[4] Joseph L Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin* 76, 5 (1971), 378.

[5] Bo Geng, Linjun Yang, Chao Xu, Xian-Sheng Hua, and Shipeng Li. 2011. The Role of Attractiveness in Web Image Search. In *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 63–72.

[6] Sven Ove Hansson and Till Grüne-Yanoff. 2008. Preferences. *Stanford Encyclopedia of Philosophy* (2008).

[7] Kai Hui and Klaus Berberich. 2017. Low-cost Preference Judgment via Ties. In *European Conference on Information Retrieval*. Springer, 626–632.

[8] Kai Hui and Klaus Berberich. 2017. Transitivity, Time consumption, and Quality of Preference Judgments in Crowdsourcing. In *European Conference on Information Retrieval*. Springer, 239–251.

[9] Vidit Jain and Manik Varma. 2011. Learning to Re-rank: Query-dependent Image Re-ranking Using Click Data. In *Proceedings of the 20th International Conference on World Wide Web*. ACM, 277–286.

[10] Kalervo Järvelin, Kalervo Järvelin, and Jaana Kekäläinen. 2000. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 41–48.

[11] Gabriella Kazai, Emine Yilmaz, Nick Craswell, and Seyed MM Tahaghoghi. 2013. User Intent and Assessor Disagreement in Web Search Evaluation. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 699–708.

[12] Damien Lefortier, Pavel Serdyukov, and Maarten de Rijke. 2014. Online Exploration for Detecting Shifts in Fresh Intent. In *CIKM 2014: 23rd ACM Conference on Information and Knowledge Management*. ACM.

[13] Alistair Moffat and Justin Zobel. 2008. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), Article 2.

[14] 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*. ACM, 559–568.

[15] Jose L Pardo-Vazquez, Juan R Castiñeiras-de Saa, Mafalda Valente, Iris Damião, Tiago Costa, M Inês Vicente, André G Mendonça, Zachary F Mainen, and Alfonso Renart. 2019. The Mechanistic Foundation of Weber's Law. *Nature Neuroscience* (2019), 1–10.

[16] Kira Radinsky and Nir Ailon. 2011. Ranking from Pairs and Triplets: Information Quality, Evaluation Methods and Query Complexity. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining*. ACM, 105–114.

[17] Mark E Rorvig. 1990. The Simple Scalability of Documents. *Journal of the American Society for Information Science* 41, 8 (1990), 590–598.

[18] Jitao Sang, Changsheng Xu, and Dongyuan Lu. 2011. Learn to Personalized Image Search from the Photo Sharing Websites. *IEEE Transactions on Multimedia* 14, 4 (2011), 963–974.

[19] Yunqiu Shao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2019. On Annotation Methodologies for Image Search Evaluation. *ACM Transactions on Information Systems (TOIS)* 37, 3 (2019), Article 29.

[20] Ruihua Song, Qingwei Guo, Ruochi Zhang, Guomao Xin, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2011. Select-the-Best-Ones: A New Way to Judge Relative Relevance. *Information Processing & Management* 47, 1 (2011), 37–52.

[21] Ernst Heinrich Weber. 1834. *De Pulsu, Resorptione, Auditu et Tactu: Annotationes Anatomicae et Physiologicae, Auctore*. prostat apud CF Koehler.

[22] 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*. ACM, 645–653.

[23] 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*. ACM, 655–663.

[24] 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*. ACM, 275–284.

[25] Xiaohui Xie, Jiaxin Mao, Maarten de Rijke, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2018. Constructing an Interaction Behavior Model for Web Image Search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 425–434.

[26] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Qingyao Ai, Yufei Huang, Min Zhang, and Shaoping Ma. 2019. Improving Web Image Search with Contextual Information. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, 1683–1692.

[27] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Yunqiu Shao, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Grid-based Evaluation Metrics for Web Image Search. In *The World Wide Web Conference*. ACM, 2103–2114.

[28] Xiaopeng Yang, Yongdong Zhang, Ting Yao, Chong-Wah Ngo, and Tao Mei. 2015. Click-boosting Multi-modality Graph-based Reranking for Image Search. *Multimedia Systems* 21, 2 (2015), 217–227.

[29] Y.Y. Yao. 1995. Measuring Retrieval Effectiveness Based on User Preference of Documents. *Journal of the American Society for Information science* 46, 2 (1995), 133–145.

[30] 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.