Does Diversity Affect User Satisfaction in Image Search

ZHIJING WU, Tsinghua University, China
KE ZHOU, University of Nottingham & Nokia Bell Labs, United Kingdom
YIQUN LIU, MIN ZHANG, and SHAOPING MA, Tsinghua University, China

Diversity has been taken into consideration by existing Web image search engines in ranking search results. However, there is no thorough investigation of how diversity affects user satisfaction in image search. In this article, we address the following questions: (1) How do different factors, such as content and visual presentations, affect users' perception of diversity? (2) How does search result diversity affect user satisfaction with different search intents? To answer those questions, we conduct a set of laboratory user studies to collect users' perceived diversity annotations and search satisfaction. We find that the existence of nearly duplicated image results has the largest impact on users' perceived diversity, followed by the similarity in content and visual presentations. Besides these findings, we also investigate the relationship between diversity and satisfaction in image search. Specifically, we find that users' preference for diversity varies across different search intents. When users want to collect information or save images for further usage (the *Locate* search tasks), more diversified result lists lead to higher satisfaction levels. The insights may help commercial image search engines to design better result ranking strategies and evaluation metrics.

CCS Concepts: • **Information systems → Users and interactive retrieval**; *Web search engines*;

Additional Key Words and Phrases: Image search, image diversity, user satisfaction

ACM Reference format:
Zhijing Wu, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Does Diversity Affect User Satisfaction in Image Search. *ACM Trans. Inf. Syst.* 37, 3, Article 35 (May 2019), 30 pages.
https://doi.org/10.1145/3320118

1 INTRODUCTION

Search diversification is considered as an effective way to solve the problems of query ambiguity and result redundancy in Web search [5]. In recent years, many research studies focus on improving the diversity of top-ranked search results including extrinsic (topic coverage-based) and intrinsic (novelty-based) diversified ranking strategies [35]. The former comes from ambiguity in the entity the query refers to (e.g., the query "puma" can refer to either a cat or a sportswear brand) or uncertainty about the user (e.g., for query "swine flu," doctors and patients may be interested in different aspects). The latter intrinsic diversity considers information redundancy and aims at

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).
Authors' addresses: Z. Wu, Y. Liu (corresponding author), M. Zhang, S. Ma, Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China; emails: wuzhijing.joyce@gmail.com, {yiqunliu, z-m}@tsinghua.edu.cn, msp@mail.tsinghua.edu.cn; K. Zhou, University of Nottingham & Nokia Bell Labs, United Kingdom; email: zhouke.nlp@gmail.com.

1046-8188/2019/05-ART35 $15.00
https://doi.org/10.1145/3320118

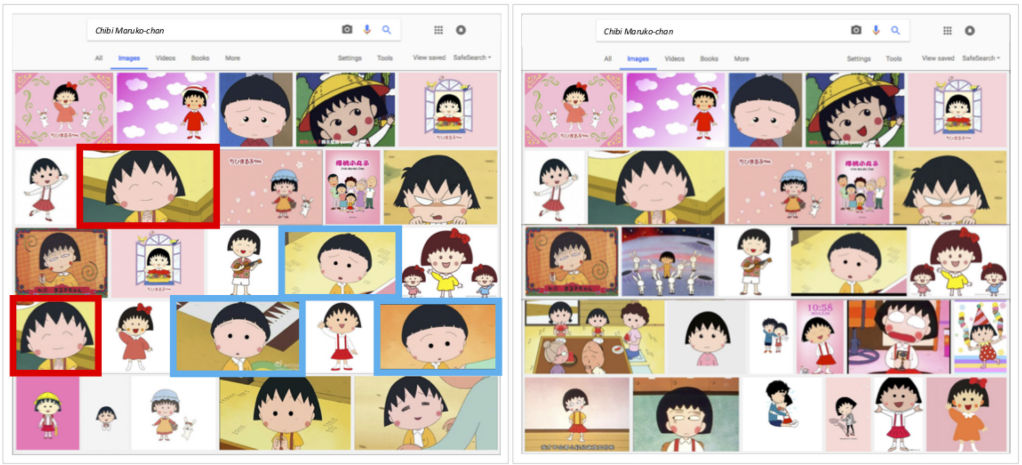


Fig. 1. Two image search result pages of query "Chibi Maruko-chan" (a Japanese cartoon character) that vary on diversity [best viewed in color]. Red boxes mark near-duplicate results (two parts of the same image), and blue boxes mark the similar results (similar in color, composition, and expression).

presenting novel and useful results. Various prior research studies investigate how to exploit the textual contents, such as the textual terms [5] and entities [41], to mine subtopics (or intents, aspects, facets) of the query to diversify the search results.

Different from the general web search engine that retrieves text-based web pages, image search returns visual images to the users. Similarly, for ambiguous or multi-faceted information needs on image search, diversification may potentially help users to explore the information spaces more effectively. Several research papers diversify the image search results by mining the similarity of visual cues on the images [11, 49], such as color or edge histogram. Others exploit the textual tags alongside the images [16] and diversify using similar techniques of the text-based document diversification.

Although these studies shed light on how to diversify for image search, few existing work investigates what factors affect users' perception of diversity for images. Both content (such as objects) and visual presentations (such as color) of the images may have an effect on the diversity level of search results users perceive. This motivates our first research question:

RQ1: How do different factors, such as content and visual presentations, affect users' perception of diversity in image search?

In addition, most of image diversification studies [11, 16, 49] assume that users prefer more diversified search engine result pages (SERPs) and therefore are more satisfied during the search session. To our knowledge, none of these studies have investigated the effectiveness of image search diversification from the user perspective. Figure 1 shows the top five rows of image search results of the query "Chibi Maruko-chan" (a Japanese cartoon character). We can see that the left SERP contains results that are near-duplicate or similar in color, composition, and expression. The right SERP is relatively more diversified in visual representations and content. Both SERPs provide useful results to users, while whether users indeed prefer diversified image search results to non-diversified ones remains under-investigated. With different search intents, users perceive satisfaction and interact with image search engines in different ways [54]. Some query types are associated with exploratory, browsing-style behavior, while for other query types users exhibit

a more focused search. However, whether users prefer diversified results in image search across different search scenarios has not been investigated. This motivates our second research question:

RQ2: How does search result diversity affect user satisfaction in image search?

In this study, we demonstrate how different factors affect users' perception of diversity and how diversity affects user satisfaction in image search. We first set experimental search tasks that cover the primary search intents in image search and then generate candidate SERPs that vary on diversity for analysis. To answer **RQ1**, we employ annotators to provide diversity annotations and collect the influencing factors of their annotations through an interview. Based on the insights gathered during the interview, we divide these prime factors into three categories: (near) duplicate images, visual presentation, and content. We find that whether there exist (near) duplicated image results has the most significant impact on users' diversity perception, followed by the similarity in content and then that in visual presentations. Furthermore, we conduct another two annotation experiments on users' perception of visual and content diversity. The substantial assessment consistency demonstrates that even users use different criteria in defining these two types of diversity, the perceived diversity on whether the SERP is diverse can be relatively similar. We then set out to answer **RQ2**, which attempts to demonstrate how diversity affects user satisfaction for a variety of different image search intents. Through a laboratory user study, we find that users' preference for diversity varies across different search scenarios. When users want to collect information or save images for further usage (the *Locate* intent type), more diversified result lists lead to higher satisfaction levels.

To summarize, our main contributions[1] are as follows:

- Through an interview, we analyze in-depth a variety of factors that can affect users' perceived diversity in image search;
- We establish the relationship between diversity and user satisfaction in image search for the first time. Indeed, a more diversified search result page can result in higher user satisfaction for certain types of tasks.

The organization of the article is as follows: In the next section, we review related work. We provide an overview of our work in understanding image search diversity in Section 3. Section 4 introduces our user study search task design. We report the analysis results of diversity annotation and user satisfaction feedback in Sections 5 and 6, respectively. Finally, we discuss our results, implications, and limitations in Section 7 and conclude the article in Section 8.

2 RELATED WORK

Three lines of research work are related to our research questions: search result diversification, search user satisfaction, and user behavior in image search.

2.1 Search Result Diversification

Search result diversification has received a lot of attention in recent years. It can be taken as a task of representative set covering problem to ensure good coverage of users' information needs. One popular solution for diversification is that all the relevant results are generated at first with a standard retrieval model. After that, they are re-ranked using a clustering method to find a subset with high diversity [3, 11, 56]. Another method is using a greedy algorithm to directly retrieve points sequentially by combining the relevance and diversity scores [37, 44].

[1]The data are available at https://drive.google.com/open?id=139zWD4oBlzWXc3vzjk4NQ1-ykwPzyUHI.

The seminal work on diversity in information retrieval is the use of maximal marginal relevance (MMR) [5]. Carbonell proposed MMR to linearly combine independent measurements of relevance and diversity to reduce redundancy and then used it for re-ranking documents and producing summaries. In their approach, there is no categorization of the document or query, diversification is conducted through the choice of similarity functions. Then Das Sarma et al. [10] solved a similar problem of finding sets of results that are unlikely to be collectively bypassed by a typical user and designed a greedy approach to achieve this objective. Another similar greedy algorithm is developed for maximizing the probability of finding at least one relevant document in top-ranked results. It is demonstrated to promote diversity by looking at ambiguous queries [7]. Agrawal et al. [1] presented a systematic approach to diversifying results that aims to minimize the risk of dissatisfaction of the average user. Different from Carbonell and Goldstein [5], they considered cases in which both queries and documents may belong to more than one category according to a taxonomy of information. Radlinski and Dumais [36] presented and evaluated methods for diversifying search results to improve personalized web search. Various techniques have been proposed to diversify search results by explicitly modeling the search intents underlying the query and documents [39–41].

With respect to Web image search, extensive efforts are dedicated to making the top-ranked image results diversified. Several IR (Information Retrieval) challenges are organized to promote the diversity of retrieved images, such as the *Retrieving Diverse Social Images* task[2] of *MediaEval* and the *ImageCLEF Photo Retrieval* task.[3] Both tasks combine the relevance and diversity of search results for evaluation. The F-1 score is used to calculate the harmonic mean of relevance-based precision (the fraction of relevant documents) and cluster-based recall (the proportion of subtopics retrieved in documents) [25]. When calculating the similarity between two images, most existing image similarity models calculate the category-level image similarity based on the content of an image [14, 51]. Other methods are based on image features to calculate the visual similarity [6, 9, 27, 58]. Wang et al. [50] converted each image into a K-bit hash code according to its content and proposed a fast and effective detection algorithm to detect all visually duplicate groups in large image collections. Tong et al. [47] provided a measure to quantify goodness for a top-k ranking list that captures both relevance and diversity and proposed an algorithm to find a diversified top-k ranking list from large graphs. Similarly, Yan et al. [56] considered the problem of clustering and re-ranking Web image search results so as to improve diversity at high ranks. Furthermore, a supervised version of the popular MMR diversification algorithm was proposed to improve both the relevance and the diversity of the top results in image search [44]. Different from the MMR algorithm, hash functions were used to characterize the locality sensitive hashing to retrieve approximate nearest neighbors in sub-linear time with superior diversity [37]. Recently, Qian et al. [34] proposed a topic diversified ranking approach considering the topic coverage of the retrieved images. They used inter-community, and intra-community ranking methods to achieve a good tradeoff between the diversity and relevance performance.

All those existing research studies made an assumption that users prefer more diversified results. However, how to define diversity in image search scenarios and whether users indeed prefer diversified image search results to non-diversified ones remain under-investigated.

2.2 User Satisfaction

Satisfaction has been studied extensively in psychology [43], commerce [31], and some other fields. In the IR literature, search satisfaction was first introduced in the 1970s according to Su [45], and

[2] http://www.multimediaeval.org/mediaeval2017/diverseimages/.
[3] https://www.imageclef.org/.

it is generally defined as the fulfillment of a user's specified desire or goal [17]. Wang et al. [52] demonstrated the clear utility of the satisfaction labels on improving performance in document relevance estimation and query suggestion. Therefore, researchers tried to collect satisfaction feedback to improve the performance of search engine. Fox et al. [13] compared user search behavior with satisfaction and found a strong association between users' search patterns and their explicit satisfaction ratings. It has been shown that users' search behaviors provide more accurate signals of search satisfaction than query-document relevance [17].

User behaviors have been extensively used to predict user satisfaction. Guo et al. [15] showed that fine-grained interactions, such as mouse cursor movements and scrolling, provide additional clues for better predicting satisfaction of a search session as a whole. Hassan et al. [18] studied additional implicit signals based on the relationship between the user's current query and the next query, such as their textual similarity and the inter-query time. They showed that a query-based model (with no click information) can indicate satisfaction more accurately than click-based models. Similarly, a good measurement of search satisfaction was provided in the absence of clicks through studying whether tracking the browser viewport (visible portion of a web page) on mobile phones could enable accurate measurement of user attention at scale [23]. Kim et al. [22] utilized three measures of dwell time for predicting click-level satisfaction. Liu et al. [26] first attempted to predict search satisfaction with mouse movement patterns (motifs) on SERPs. They proposed to use distance-based and distribution-based strategies in the selection of motifs, which outperforms existing frequency-based strategy in choosing the most effective motifs to separate SAT sessions from DSAT ones. A recent study [29] aimed at extracting interpretable user interaction subsequences and focused on informative action sequences rather than mouse movement coordinates for predicting search satisfaction.

However, to the best of our knowledge, few existing studies take diversity into consideration when predicting user satisfaction in either general Web search or image search. We aim to establish the relationship between diversity and satisfaction in image search for the first time.

2.3 User Behavior in Image Search

Investigating user behavior in image search allows us to better understand users' image search process. With the wide application of Web search engines, the analysis of query logs becomes one of the most common approaches to understand user behavior. For general Web search, previous studies focus on the individual query, sessions, and click position on SERPs [20, 42], which helps us gain a better understanding of how users use search engines. Since the way to present results in image search differs greatly from that of general Web search, large-scale log analysis is made to investigate the different behavior between general Web and image search [2, 33, 53]. Compared to general Web search, image search users usually submit shorter query strings and their selections of query terms are more diverse. They also found that image searches lead to more clicks and exploratory behavior. Clicked images for the same query vary greatly across users. Many features such as session length, browsing depth, and query reformulation patterns are also measured to characterize the general behavior of image search users [19, 30].

Understanding what users search for and why users search provides the context for search behavior. Search intents of general Web are classified into three categories by Broder [4]: informational, navigational, and transactional. Lux et al. [28] adapted this taxonomy for image search. They categorized user intents into knowledge orientation, mental image, navigation, and transaction. Park et al. [32] analyzed a large-scale query log from Yahoo Image Search to investigate user behavior toward different query types and identified important behavioral differences across them. Xie et al. [54] showed that user intents in image search can be grouped into three classes, Explore/Learn, Entertain, and Locate/Acquire, which we follow in our search task design. Users have

different behavior patterns under these three intents, such as first click time, query reformulation, dwell time, and mouse movement on the result page.

Since examination behavior in search is closely related to users' attention distribution, it has been investigated by several researchers through eye-tracking devices. This can yield more detailed observations about how users interact with the search result. Tatler and Vincent [46] showed that understanding biases in how users move the eyes can provide powerful new insights into the decision about where to look in complex scenes. Underwood and Foulsham [48] investigated how eye movements are affected by visual saliency. Xie et al. [55] conducted a laboratory eye-tracking study for image search and found that users' examination patterns have a middle-position bias. Besides the position factor, the content of image results such as visual saliency affects examination behavior.

There are few researches that study user behavior of image search under different diversity levels. In our work, we aim to further understand how search result diversity affects image search user behavior (such as satisfaction) for a variety of image search tasks.

3 DIVERSITY IN IMAGE SEARCH

The main objectives of this work are twofold: (1) understanding what *diversity* means in the context of image search, i.e., what and how different factors affect image search users' perception of diversity, and (2) investigating how diversity affect users' satisfaction for a variety of image search intents.

We start by stating the topic of diversity we focused on throughout the article. In a general Web search, diversity is summed up as two prime types: extrinsic diversity and intrinsic diversity [35].

- *Extrinsic diversity*: aims to address uncertainty about the information needs given a query.
- *Intrinsic diversity*: aims to avoid redundancy and thus present a novel and useful set of results under a single well-defined information need.

Uncertainty about the information need (*extrinsic diversity*) may be caused by different users or query ambiguity. For example, the potential intents for query "Apple" include "fruit apple," "devices from Apple," "Apple services," and so on. We do not know whether the user wants to buy a new iPhone or gather information for different varieties of apples. The latter *intrinsic diversity* aims to find results that cover different aspects of a specific information need and optimize for novelty in results. This type of diversity is required when the information need cannot be satisfied by a single result or user desires a selection of options to choose between. Potentially, the user is seeking different views or results from different sources to build confidence in the correctness of the answer to her information need. Similarly, take the case of query "Apple"; if a user's detailed search intent is to buy a new iPhone, she may want to get a variety of reviews about iPhone models and compare prices from different shopping websites.

In this article, when we study *diversity* in the context of image search, we focus on *intrinsic diversity* rather than extrinsic diversity. The information needs for users are designed to be specific, with clearly defined search tasks to accomplish. We are more interested in those scenarios and aim to understand whether diversity may still make a difference in user satisfaction. An overview of our work is shown in Figure 2 and can be summarized as follows:

- Preliminaries for diversity analysis: We set 20 search tasks with scenario description to ensure that the information needs are specific. We manipulate the search results for each task and control the placement of similar image results. Since our focus in this step is not to propose an algorithm to automatically generate SERPs that vary in diversity, we use the simple SIFT algorithm to calculate the similarity between two images and generate SERPs

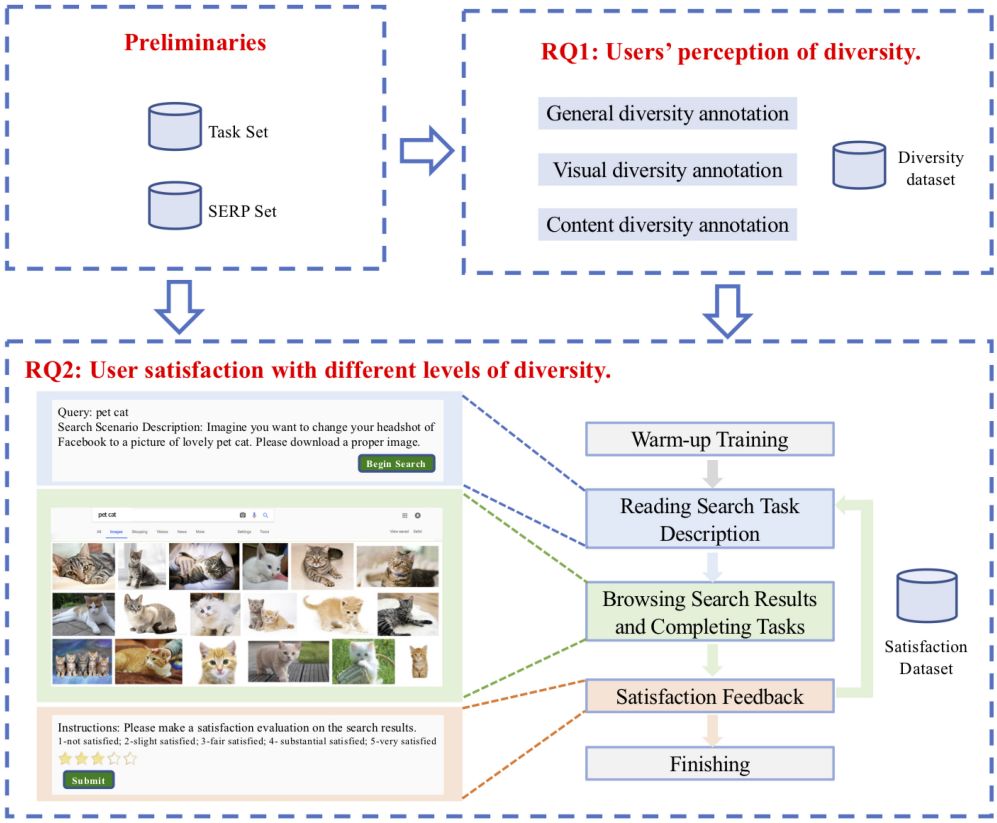


Fig. 2. Summary of our experiments in this article.

that vary in diversity. These candidate SERPs are used for diversity annotation experiments to facilitate diversity analysis (see Section 4).

- Diversity in image search: As noted above, users may perceive diversity in general Web search according to the website, sub-topic, and query-dependent factors such as phone models and prices. Since the search scenario in image search differs from that in general Web search, we begin by investigating what factors affect users' perception of intrinsic diversity in image search. We employ the assessors to annotate their perceived diversity of these candidate SERPs and then ask them the underlying reasons for their decisions through an interview. By using an open-coded discussion methodology [38], we identify the main factor categories that affect user perceived diversity in image search: near duplicate, visual diversity, and diversity in content (see Section 5.1). The detailed factors and image examples of those factor categories are shown respectively in Table 8 and Figure 6. Given that both visual and content diversity are seemed to be important by users on their perceived diversity, we further conduct two annotation experiments to understand how users perceive these two types of diversity (see Section 5.2).
- Diversity vs. satisfaction: Satisfaction is defined as the fulfillment of information requirement [21], which measures users' subjective feelings during the search processes (i.e., in accomplishing the search task in our context). The study of the relationship between diversity and user satisfaction (collected from explicit user feedback) in Section 6 is based on both diversity types/annotations: visual and content. We are interested in finding out how those

different types of intrinsic diversity affect user satisfaction and whether the relationship differ according to different image search intents (*locate* vs. *learn* vs. *play*, see Section 4).

4 EXPERIMENTAL DESIGN

To analyze the influencing factors of perceived diversity and further establish its relationship to user satisfaction, we start by setting experimental search tasks that cover the primary search intents in image search and generating candidate SERPs of these search tasks that vary on diversity.

4.1 Search Task Design

To set our search tasks, we follow the existing search intent taxonomy from previous work on image search [54]. Distribution of each intent is investigated through analysis of a commercial Web image search log. We finally design 20 image search tasks based on the real-world intent distribution.

4.1.1 Search Intent Taxonomy. To capture various popular image search intents and subsequently investigate whether diversity plays different roles for those different intents, we need to choose an appropriate image search intent taxonomy. Due to the lack of previous work that studies image diversity under different search intents, we aim to take the first step by following the most recent (and basic) intent taxonomy for image search proposed by Xie et al. [54]. According to the criterion of "Is the user's search behavior driven by a clear objective?," they divide user intent into two groups. In some cases, the user freely browses the image search results for entertainment without a clear objective. In other cases, they consider the criterion of "Does the user need to download the image for further use after the search process?" For some search tasks, people have to download images for further use, whereas for other tasks, people are capable of satisfying their information need without downloading images. According to these criteria, we adopt the following image search intent taxonomy:

- **Locate:** The user is looking to download something for further use. Example: finding a landscape image as computer wallpaper ("locate/acquire" from Reference [54]).
- **Learn:** The user is looking to discover something or learn about a topic or confirm or compare information by browsing images. They can obtain, check, or compare information by examining images in result pages only. Example: browsing and compare some images of different decoration style to find the most proper one for your house ("explore/learn" from Reference [54]).
- **Play:** The user just wants to browse images for fun to kill time. Example: browsing some pictures of your favorite movie stars or some humorous images in your leisure time ("entertain" from Reference [54]).

4.1.2 Intent Distribution in Search Logs. To set the proportion of each search intent, we asked three annotators to manually annotate the search sessions into the intent categories (*Locate*, *Learn*, *Play*). We sample 200 search sessions in March 2017 from the search logs of a popular commercial image search engine. A session contains consecutive queries issued by a single user within a short time period [42]. We partition a user's search actions into separate sessions when the time between consecutive actions exceeds 30 minutes [32]. For each session, we extract the query list, browsing depth, and clicks. Annotators are shown this information and asked to give annotation (1-*Locate*, 2-*Learn*, 3-*Play*, 4-*Others*). *Others* refers to those sessions that cannot be directly mapped to these three categories. We report the value of Fleiss' Kappa [12], a statistical measure for assessing the consistency/reliability of annotator agreement. According to Landis and Koch [24], Fleiss' Kappa ranges from 0 to 1 (0–0.2: slight agreement; 0.2–0.4: fair agreement; 0.4–0.6: moderate agreement; 0.6–0.8: substantial agreement; 0.8–1.0: almost perfect agreement). The Fleiss' Kappa is 0.79 among

Table 1. "Locate" Tasks Used in Our Study

ID	Query	Search Scenario Description
1	praying	Imagine you want to find some pictures about people praying and add them to your presentation slides.
2	Wu Yifan	Imagine you like the Chinese actor "Wu Yifan" and you want to find some photos of him for the computer wallpaper.
3	cartoon mouse	Imagine you want to change your Facebook profile picture to a picture of a lovely cartoon mouse. Please find and download some appropriate images.
4	Chibi Maruko-chan	Imagine you want to find some pictures of the Japanese cartoon character "Chibi Maruko-chan" as your social-networking WeChat app profile picture.
5	abdominal muscles	Imagine you want to download to your mobile phone some photos of people who are fit and having abdominal muscles to motivate yourself to lose weight.
6	basketball	Imagine you are making some presentation slides that introduces basketball. You want to find some basketball-related pictures and add them to it.
7	slides background	Imagine you are making a slide and you want to find some theme background pictures for it.
8	Titanic	Imagine you are making a slide that introduces the classic movies Titanic. You want to find some relevant pictures of Titanic.
9	cartoon character	Imagine you want to find a picture of a cartoon character as your social-networking WeChat app profile picture.
10	landscape	Imagine you want to find a landscape picture as the cover photo of your Facebook profile.

three annotators that leads to a substantial agreement [24]; 96.5% of sessions can be classified into *Locate*, *Learn*, or *Play*, and the proportions of those categories are respectively 47%, 23.5%, and 26%. This demonstrates that these three categories can cover most of the users' search intents within image search. This thus also verifies our choice on the taxonomy.

4.1.3 Search Tasks. We design 20 search tasks based on the intent distribution above (10 belong to *Locate* intent, 5 belong to *Learn* intent, and 5 belong to *Play* intent). The tasks used in our study are shown in Tables 1, 2, and 3. A search scenario description is provided to avoid possible ambiguity in the query text and ensure each participant faces the same task difficulty level. For instance, in the query "decoration styles" of *Learn* tasks, participants are asked to browse images of different decoration style for reference. The information need can be satisfied by only examining images in SERPs. For the query "cartoon mouse" of *Locate* tasks, we not only ask participants to browse and find proper images of a cartoon mouse but also ask them to choose one and download it. They need to consider carefully when choosing images for further use from the image results presented on the SERP. On the contrary, we do not set a specific search goal for the query "short haircuts" of *Play* intent. Participants can feel free and just look around to kill time. All or our further analyses in this article are based on these search tasks.

4.2 SERP Generation

In this section, we generate two candidate SERPs that vary on diversity (*non-diversified* and *diversified*) for each search task. To obtain a set of image search results, we issue the 20 task queries

Table 2. "Learn" Tasks Used in Our Study

ID	Query	Search Scenario Description
1	voucher	Imagine you are designing vouchers for your friend's shop. You want to browse some pictures of vouchers for reference.
2	decoration style	Imagine you just bought a house and plan to decorate it. You want to browse some images of different potential decoration styles for reference.
3	Hebei province	Imagine you are going to travel to Hebei province of China during your holidays. You want to look for maps of Hebei to find out some points of interest to visit.
4	porcelain	Imagine you have just visited a museum where you saw many beautiful porcelains. You want to look for more pictures of them and observe their patterns.
5	edible bird's nest	Edible bird's nest is considered as a nutritious food that is good for health in China. Imagine you want to see how it looks like before cooking.

Table 3. "Play" Tasks Used in Our Study

ID	Query	Search Scenario Description
1	Hu Ge and Jiang Shuying	Imagine you are one of fans of two Chinese actors Hu Ge and Jiang Shuying, and wish them to be married in the future. You want to casually browse some group photos of them.
2	still photos of "Novoland: The Castle in the Sky"	Imagine you have watched the Chinese television series "Novoland: The Castle in the Sky" while you think some of the scenes in the series are beautiful. You want to appreciate and view more still photographs in your spare time.
3	young Liu Dehua	Imagine you like the Chinese actor and singer Liu Dehua very much. You want to casually browse some photos of him when he was young.
4	G-Dragon's girlfriend	Imagine you hear about the news that the Korean star "G-Dragon" has a girlfriend now and you want to take a look at her photos to kill time.
5	short haircuts	Imagine you are in a boring lecture. You are looking around on the Internet and incidentally browsing some images of short haircut styles to kill time.

into a commercial image search engine and collect the top 100 retrieved image results per query. Specifically, we organize the collected top retrieved images in rows in our experimental system. To avoid any potential bias brought by the varying number of image results to users' perception, we reserve five rows of image results for each SERP.

Figure 1 shows the examples of our generated *non-diversified* and *diversified* SERPs for one of our search tasks "Chibi Maruko-chan," the cartoon character. We treat the original ranking of image results returned by the commercial image search engine (left SERP in Figure 1) as *non-diversified* SERP. However, there exist (near) duplicate image results (marked by the red boxes). The image in the fourth row can be seen as part of the image in the second row. There are also

ALGORITHM 1: A heuristic algorithm to generate *diversified* SERPs.

Input: I: a list of top 100 retrieved image results;
$count$: quantity of image results that placed in top 5 rows;
θ: similarity threshold manually set on the query-basis;
$f(D)$: return the minimum k that satisfies $Similarity(I_k, I_i) \leqslant \theta$ for each $I_i \in D$ ($k > count$);
(Experiment results show that $f(D)$ can always return a valid k in our dataset.)
Output: D: a list of top 5 rows of image results for diversified SERP
$D = [I_1]$;
for $i = 2$ *to* $count$ **do**
  for I_j *in* D **do**
    $Similarity(I_i, I_j) = SIFT(I_i, I_j)$;
    if $Similarity(I_i, I_j) > \theta$ **then**
      $k = f(D)$;
      $I_i = I_k$;
      add I_i to D;
      break;
    end
    if $j == i - 1$ **then**
      add I_i to D;
    end
  end
end

several similar image results (marked by the blue boxes). The subjects of those images have the same expression and clothes. By utilizing the top 100 retrieved image search engine results, we generate the *diversified* SERP by a heuristic algorithm described in Algorithm 1.

The basic idea of this heuristic algorithm is to find those images that are similar to previously presented image search results and then replace them with other dissimilar images. We use I to represent the list of image results and *count* to represent the number of image results that placed in the top five rows (the number of image results we reserve for *non-diversified* SERP). Through this heuristic process, we aim to get a set of image results for *diversified* SERP, the size of which is *count*. For each image I_i in I, we use the Scale-invariant feature transform (SIFT[4]) [27] algorithm to calculate the similarity between I_i and $I_j (0 < j < i)$ and then manually set the similarity threshold θ on the query-basis. The value manually chosen for θ ranged from 0.09 for query 7 to 0.47 for query 13, with an average value of 0.24. If I_i is similar with any I_j ($Similarity(I_i, I_j) > \theta$), then we replace I_i with I_k, an image that is ranked further than the top five rows and dissimilar with each I_j. If the sizes of I_i and I_k are not the same, then I_k is scaled to ensure that both the width and height are equal to or greater than that of I_i. We then manually crop I_k to ensure that I_k and I_i are of the exactly equal size. In this work, we do not exploit more advanced methods, such as the clustering methods proposed in the *ImageCLEF* task to manipulate the diversity of SERPs. This is because our focus is not to propose the most accurate SERP manipulation approaches; how to appropriately cluster query-specific image results and then diversify is still an open research question. For our purpose, although our proposed heuristic method is simple, it has been demonstrated to be effective on generating candidate SERPs (see Section 5.1 and Table 7 for more details on the evaluation).

[4]Scale-invariant feature transform (SIFT) is a state-of-the-art visual algorithm that extracts distinctive invariant features and generates a large number of features that densely cover the image over the full range of scales and locations. http://www.cs.ubc.ca/~lowe/keypoints/.

Table 4. Comparison of State-of-the-Art Offline Image Search Evaluation Metrics Based on Topical Relevance between the Generated *Diversified* and *Non-diversified* SERPs

	Row-AVG			T-sequence		
	CG	DCG@5r	RBP(0.95)	CG	DCG@5r	RBP(0.95)
Non-diversified SERP	4.72	2.81	0.21	24.0	7.81	0.69
Diversified SERP	4.47	2.66	0.20	22.9	7.41	0.66

Two set of metrics (Row-AVG and T-sequence) are adopted, since they correlate best with user satisfaction according to Zhang et al. [57]. We observe that the differences of these metrics between *diversified* and *non-diversified* SERPs are all not statistically significant (T-test, p-value > 0.5).

4.3 Analysis of Generated SERPs

Since our heuristic algorithm replaces some top-ranked images with images positioned further in the ranking, the relevance of search results may change after this diversification process. To further analyze the potential differences, we employ three external experts (who are all familiar with the relevance annotation task) to make topical relevance assessments on the images. Each assessor is required to provide relevance annotations for all query-image pairs that have been presented in the 20 *non-diversified* and 20 *diversified* SERPs. A simple 2-point topical relevance scale is used for the relevance annotation:

- **Relevant:** The main subject of the images is the subject of the query and is clearly visible. If there is more than one subject in the query, then the images should contain all subjects.
- **Non-relevant:** The image fails to match the subject of the query or just matches part of subjects of the query.

The Fleiss' Kappa among the three annotators is 0.637, which leads to a substantial agreement. We use the majority vote of three assessors as the relevance label of a query-image pair, which means that at least two annotators annotate the given query-image pair with the same relevance level. Given the relevance annotations, we aim to use the state-of-the-art offline evaluation metrics for image search to evaluate the relevance of the image SERPs (both diversified and non-diversified). According to the meta-evaluation results reported by Zhang et al. [57], it has been shown that two sets of offline metrics, "Row-AVG" and "T-sequence," correlate best with user satisfaction in the context of image search. The "Row-AVG" method takes image results in a row as an integrated result with the average of topical relevance for the images in the row. The "T-sequence" method considers the middle-position bias of user's examination behavior in image search [55] and obtains a "T" sequence from a two-dimensional results placement. Given the considered examination biases in those two approaches, the two-dimensional image search metrics can be calculated like traditional metrics in the one-dimensional ranking list-based SERPs. Therefore, in this work, we adopt those two state-of-the-art image search metrics to evaluate SERP relevance and compute several widely used metrics including CG, DCG, and RBP.

As shown in Table 4, we can observe that non-diversified SERPs perform slightly better than diversified SERPs on these metrics on average, while the differences between them are not statistically significant according to T-test. To further analyze this, we plot the performance on Row-AVG metrics of the generated *diversified* and *non-diversified* SERP of each search task in Figure 3. We find that for most of the search tasks, diversified and non-diversified SERPs have very similar evaluation metric scores while for almost all tasks, non-diversified SERPs outperform or is comparable to diversified SERPs in terms of relevance. The most significant differences can be observed for the 5th (query: abdominal muscles) and 13th (query: Hebei province) search tasks. This demonstrates that although the relevance differences are marginal, due to the fact that some top-ranked images

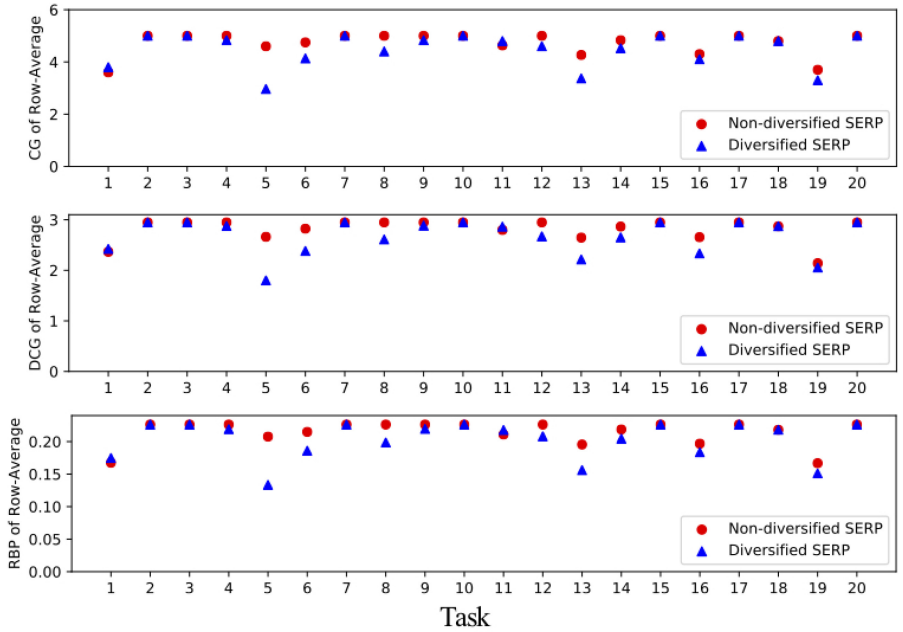


Fig. 3. Row-AVG metric performance comparison between generated *diversified* and *non-diversified* SERP for all 20 search tasks employed in our user study. For most of the search tasks, diversified and non-diversified SERPs have very similar evaluation metric scores while for almost all tasks, non-diversified SERPs outperform or is comparable to diversified SERPs in terms of relevance. The most significant differences can be observed for the 5th (query: abdominal muscles) and 13th (query: Hebei province) search tasks.

that are relevant to the query are eliminated, because they are too similar to a former image, our heuristic-based algorithm is likely to promote several non-relevant images that are ranked in the second page by the search engine.

In conclusion, we design 20 image search tasks and for each search task, we obtain both candidate *non-diversified* and *diversified* SERPs for further diversity assessments (Section 5). For most of the search tasks, both candidate *non-diversified* and *diversified* SERPs perform very similarly on topical relevance-based evaluation metrics. For a few search tasks, non-diversified SERPs outperform those diversified SERPs in terms of relevance. Note that our focus in this section is not to propose an effective and advanced algorithm that can automatically generate SERPs that vary in diversity. Rather, we utilize this heuristic algorithm to facilitate the annotations to answer our research questions. Although simple, empirically, we find this heuristic algorithm performs well in generating those *diversified* SERPs (see Section 5 and Table 7).

5 USERS' PERCEPTION OF DIVERSITY

In this section, we aim to further understand users' perception of diversity within image search to answer **RQ1**. There are many different dimensions that may affect users' perception of diversity. First, to uncover the influencing factors (Section 5.1), we employ the assessors to annotate their general perceived diversity of SERPs (generated in Section 4) and then ask them the underlying reasons of their decisions through an interview. By using an open-coded discussion methodology [38], we identify the factor categories and their importance.

Second, given that both *visual presentations* and *content* are two important factors that affect users' perceived diversity, we aim to further understand how users perceive diversity from each of

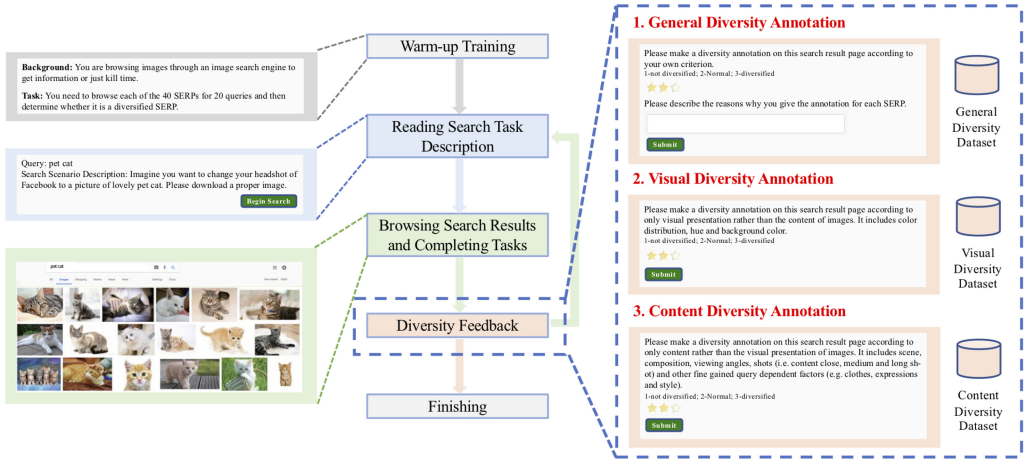


Fig. 4. Summary of three diversity annotation experiments we collect to understand image search diversity. General diversity annotation aims to identify the important factor categories that affect user perceived diversity; while both visual and content diversity annotations, the two influential diversity factors, are collected further to understand their relationships. Each type of diversity is annotated by 10 different assessors, respectively.

Table 5. Summary of All the Annotation Data Used in This Work

Unit	Assessments	Annotated by	Scales	#Annotators
Image result	Topical Relevance	Experts	2	3
SERP	General Diversity	Assessor Group 1	3	10
	Visual Diversity	Assessor Group 2	3	10
	Content Diversity	Assessor Group 3	3	10
	Satisfaction	Search User (participant)	5	30

There are no overlaps among different assessor groups (i.e., the assessors are different). All the search user study participants who provide their satisfaction assessments do not engage to provide any other assessments.

those two aspects by collecting annotations on visual diversity and content diversity, respectively (Section 5.2). An overview of how we collect the annotations is shown in Figure 4. The summary of all annotation data is described in Table 5.

5.1 Analysis of Influencing Factors

5.1.1 Data Collection. We employ 10 assessors (undergraduates majoring in humanities and social science, engineering, or arts) to make annotations on the perceived diversity of SERPs according to their own criteria. They are all reported as frequent users of Web image search. It takes about 30 minutes to finish the annotation tasks. Figure 5 shows the task description and instruction. Each participant needs to provide 3-point scaled diversity annotations to all of the 40 SERPs we generate in Section 4. To avoid direct comparisons between two SERPs of the same task, all of the 40 SERPs are shown to assessors in a random order. Although the assessors may observe both diversified and non-diversified SERPs for the same search task during the annotation, due to the randomization, the assessor rarely encounter the two SERPs of the same task subsequently. Meanwhile, we do not provide a standard definition of diversity to avoid biasing the judgments. Rather, we instruct assessors to judge by their own criteria.

Annotation Instructions 1:

Background: You are browsing images through an image search engine to get information or just kill time.

Task: You need to browse each of the 40 SERPs for 20 queries below and then determine whether it is a diversified SERP according to your own criterion. After that, you need to describe the reasons why you give the annotation for each SERP.

The diversity of SERPs can be classified as follows:
- **1 star**: Not diversified.
- **2 stars**: Normal.
- **3 stars**: Diversified.

Fig. 5. General diversity annotation instructions.

Table 6. Consistency among User's Perception of General, Content, and Visual Diversity

Intent	Fleiss' Kappa among 10 annotators		
	General diversity	Visual diversity	Content diversity
Locate	0.365	0.375	0.347
Learn	0.188	0.271	0.078
Play	0.391	0.465	0.346
All	0.333	0.373	0.298

After finishing all the annotation tasks, we ask the annotator to describe the detailed underlying reasons through an interview for each of the SERPs he/she has assessed. For instance, one of the participants provided *1 star* label (i.e., *not diversified*) to the left SERP of Figure 1. When asked about the underlying influencing factors, he answered as follows: "Several images in the SERP look almost identical. About half of the results are headshots with nearly the same posture and expression. They are similar in composition and the background contains only a flat color. Therefore I think that the SERP results are not diverse." We record these detailed answers/descriptions for our factor analysis (Sections 5.1.2 and 5.1.3).

We report the value of Fleiss' Kappa as the first column (general diversity) of Table 6 shows. The consistency among 10 annotators under all tasks is 0.333, which leads to fair agreement. This demonstrates that the perception of diversity slightly varies across different users. Users potentially have different criteria in defining diversity and therefore perceive differently on whether a SERP is diverse or not. Under the "learn" tasks, the annotators even only reach a slight agreement (the Fleiss' Kappa is 0.188). Given their specified reasons, we find that users consider more different factors of perceived diversity when focusing on comparing image results or obtaining new knowledge. To make the 3-point scaled diversity annotation comparable to the 2-point scaled generated diversity (based on our heuristic algorithm), we merge the *not diversified* and *normal* categories into one category. Then we use the majority vote of ten assessors as the general diversity label of a SERP, which means that at least six assessors annotate the given SERP with the same diversity score. Table 7 shows the joint distribution of perceived (assessed) general diversity and generated diversity of our heuristic SERP generation Algorithm 1 introduced in Section 4.2. We find that the generated (heuristic-based) diversity and perceived (assessed) general diversity are consistent with each other, except for one SERP. This indicates that our heuristic-based diversity SERP generation algorithm is effective for our purposes in this article.

Table 7. The Joint Distribution of Perceived General Diversity and Diversity Generated by Algorithm 1

generated (heuristic) \ perceived general	non-diversified	diversified
non-diversified	20	0
diversified	1	19

The generated (heuristic-based) diversity and perceived (assessed) general diversity are consistent with each other, except for one SERP.

Table 8. The Main Factor Categories That Affect Users' Perceived *Intrinsic Diversity* in Image Search, While Instrinsic Diversity Focuses on the Redundancy of Search Results Rather Than the Uncertainty about Information Needs

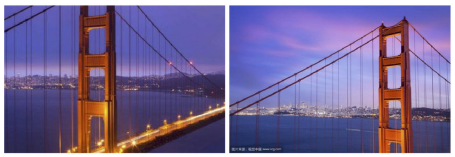
Categories	Subcategories
(Near) duplicate	(1) Whether there exist (near) duplicate image results; (2) The distance between (near) duplicate images; (3) The position of (near) duplicate image results;
Similar in visual presentation	(1) Color distribution; (2) Hue: the attribute of a color by virtue of which it is discernible as red, green, and so on; (3) Background color;
Similar in content	(1) Scene of the subjects; (2) Composition, viewing angles, shots (i.e., close, medium and long shot); (3) Fine-grained query dependent factors

5.1.2 Factor Categories. After collecting all the criteria annotators use to judge their perceived diversity through the interview, we recruit a group of three web research professionals to review all interview transcripts and discuss to determine the factor categories following Russell et al. [38]. In each iteration, the proposed factor categories were fine-tuned to cover as many factors users have mentioned as possible. Based on the discussion results, we divide the diversity factors into three broad categories and several subcategories (shown in Table 8).

When judging the diversity of a SERP, users usually pay attention to (near) duplicate and very similar images. Near-duplicate images are not really duplications. Specifically, those images could originate from different sites, have different image tags, also maintain differently in the visual components (e.g., part of the images; an example is shown in Figure 6(a)). However, those from the user point of view can be "near-duplicate" and can affect the performance. The distance and position of the (near) duplicate images can affect users' perception of diversity. For instance, it is more noticeable when two duplicate images are ranked top or presented in the viewport of the SERP (i.e., top three rows).

Visual presentation and content are another two prime aspects that affect users' perceived diversity of a SERP. The popular aspects of visual presentation that annotators mentioned are color distribution, hue, and background color. The *content* aspect includes scene, composition, viewing angles, shots, as well as some query-dependent factors such as clothes, posture, and entity style. All those factors listed in Table 8 can cover most of the influencing factors users have mentioned during the interview. We list some examples for different types of diversity in the result of query "Golden Gate Bridge" as Figure 6 shows.

(a) **Near duplicate.** These two images originate from different sites and have different image tags. However, they are part of the same image in users' perception called near duplicate results.



(b) Similar in **Visual presentation.** These two images show different parts of the bridge. However, they are similar in color distribution, hue and background color.

(c) Similar in scene of the subject (**Content**). These two images both show the bridge by sunset.

(d) Similar in composition (**Content**). These two images have different background. However, they are almost the same in composition.

Fig. 6. Example for different types of diversity in the result of query "Golden Gate Bridge."

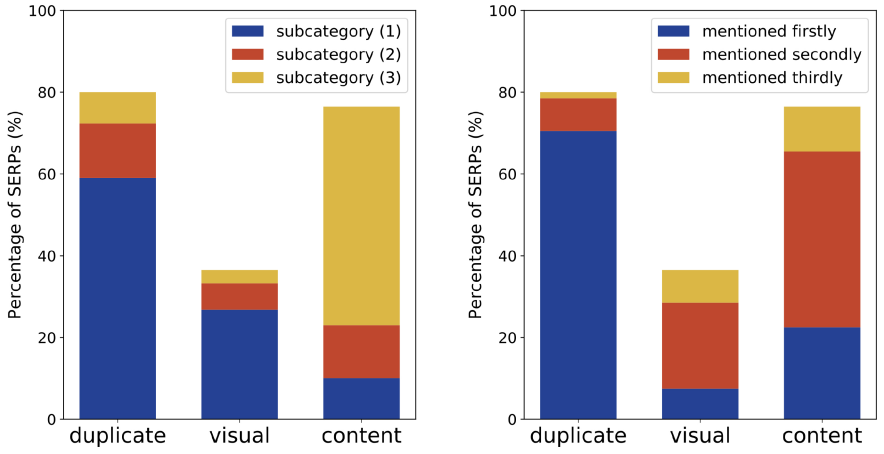


Fig. 7. The distribution of (a) subcategory; (b) order mentioned of these three factors.

5.1.3 Factor Distribution. After creating the factor categorization, we aim to investigate the importance of each factor. We identify the factors in each interview transcript we collect in Section 5.1.1 according to the factor categorization we have defined. Note that, when the users mention the factor "(near) duplicate," this does not necessarily mean there exist "(near) duplicate" images within the SERP. For example, one of the interview transcripts says: "There are no duplicate images, and the color is rich. Therefore, I think this is a diverse SERP." This just demonstrates that users take "near duplicate" into considerations when they assess the perceived diversity of the image SERP. In fact, only a few SERPs in our dataset contain (near) duplicate image results.

The distribution of subcategories and order mentioned of the three factors are shown in Figure 7. We can observe in Figure 7(a) that users have taken "(near) duplicate" into consideration for around

Annotation Instructions 2:

Background: You are browsing images through an image search engine to get information or just kill time.

Visual Diversity Annotation Task: You need to browse each of the 40 SERPs for 20 queries below and then determine whether it is a diversified SERP according to only visual presentation rather than the content of images. It includes color distribution, hue and background color.

Content Diversity Annotation Task: You need to browse each of the 40 SERPs for 20 queries below and then determine whether it is a diversified SERP according to only content rather than the visual presentation of images. It includes scene, composition, viewing angles, shots (i.e. content close, medium and long shot) and other fine gained query dependent factors (e.g. clothes, expressions and style).

The diversity of SERPs can be classified as follows:
- **1 star**: Not diversified.
- **2 stars**: Normal.
- **3 stars**: Diversified.

Fig. 8. Visual (marked by blue box) and content (marked by green box) diversity annotation instructions.

80% of SERPs. Diversity in content, however, is the second dominating factor, that was mentioned in 76.5% of SERPs. Last, users mentioned visual diversity only in 36.5% of SERPs. This implies that when assessing diversity, users pay more attention to the similarity within content than that in visual presentation. Another interesting observation is that in many cases when users deem content affects their perceived diversity, users are generally more likely to take into account query-dependent factors. From Figure 7(b), we can see that users tend to focus on whether there exist "near-duplicate" image results firstly. This indicates that whether there exist (near) duplicated image results is the dominating factor that affects users' diversity perception.

5.2 Users' Perception of Visual and Content Diversity

Although it has been found that users first focus on whether there exists (near) duplicate results when judging general diversity, only a few of SERPs in our dataset indeed contain (near) duplicate image results. For this reason, we do not further investigate this (although we provide a more in-depth discussion regarding this in Section 7) and consider another two prime diversity aspects (*visual presentations* and *content*) for the rest of the article. In this section, we conduct another two annotation experiments to further understand how users perceive visual and content-based diversity. Especially, we focus on studying (1) whether users agree with each other on their visual and content diversity perception respectively and (2) whether content and visual diversity are perceived differently.

5.2.1 Users' Perception of Visual Diversity. Different from the instruction of general diversity annotation in Section 5.1.1, we employ another 10 assessors to make annotations on the visual diversity of SERPs according to our predefined criteria (shown in Figure 8, motivated by results from Section 5.1). Based on the subcategories of visual diversity factors, we list all the possible fine-grained aspects to participants. Before the formal annotation, participants need to finish example tasks under our instructions to make sure they are familiar with the annotation procedure and understand the judging criteria. Then they are asked to provide annotations to all of the 40 SERPs we generate in Section 4.2 in random order. We report the value of Fleiss' Kappa as the

Table 9. Consistency between User's Perception of Visual and Content Diversity

Search Intent	Locate	Learn	Play	All
Fleiss' Kappa	0.699	0.6	0.583	0.650

second column (visual diversity) of Table 6 shows. The consistency among 10 annotators under all tasks is 0.373, which again leads to fair agreement. It demonstrates that even if we provide the judgment factors of diversity in our instruction, users maintain differing views and relatively low consistency. For instance, when users want to find and compare different decoration styles in the *Learn* tasks, the decoration images themselves are colorful and have no obvious entities. This would affect users' judgment on visual diversity.

5.2.2 Users' Perception of Content Diversity. Similarly to the visual diversity annotation, we employ another 10 assessors to make annotations on the content diversity of SERPs. Participants are instructed to focus on the content of the images rather than their visual presentation. The instruction is shown in Figure 8. We report the value of Fleiss' Kappa as the third column (content diversity) as Table 6 shows. Among 10 annotators, we can observe that the consistency is even lower than that in visual diversity. From a close examination of the annotations, we hypothesize this may be due to the fact that when focusing on the content of images, the similarity judgment of two images mainly depends on annotators' self-knowledge. The annotators major in different fields and have various kinds of knowledge backgrounds. They can classify the image content according to different aspects of the image content. For example, when annotators assess the search task "cartoon mouse," we find that some of them provide judgments according to the gesture diversity of the mouse while some of the others base their assessments on the style diversity. Therefore, their judgments on the content diversity have low consistency, especially on the "learn" search tasks (only 0.078). This implies that users have very different views on diversity according to the content of the images. It is even more difficult for assessors to agree on content diversity than visual diversity of the images.

5.2.3 Comparison between Visual and Content Diversity. To make a comparison between users' perception of visual and content diversity, we analyze the consistency of two annotation results. We consider the majority agreement as the final diversity category, which means that at least six annotators assign the case into the same category. As shown in Table 9, the annotations on visual and content diversity lead to a substantial agreement under all tasks as a whole. We find that in many cases, visual diversity, and content diversity are very similar to each other. For example, when dealing with the search task "landscape images," the search results contain images of mountain, river, prairie, and so on. The variance of visual features usually accords with that of content. However, the difference is more obvious under *play* tasks (moderate agreement). For example, when dealing with the search task "short haircuts," most of the annotation of visual diversity focuses on the color of haircuts. However, the annotation of content diversity mainly focuses on the shape of the haircuts or the gender of the models.

5.3 Summary

In this section, we conduct a set of annotation experiments to answer **RQ1**. Through factor analysis of interview transcripts from the general diversity annotation experiments, we find that (1) "near-duplicate images" is the dominating factor in affecting users' perceived diversity, followed by content and visual presentation, and (2) fine-grained query dependent content-based features are the second most important influencing factor for diversity perception.

Table 10. Mean Satisfaction Score Distribution of All Participants

Mean Score	[3,3.4)	[3.4,3.8)	[3.8,4.2)	[4.2,4.35]
Number of Participants	14	11	3	2

By collecting respectively visual and content diversity assessments, we find that (1) the annotations on visual and content diversity can be similar to each other, leading to a substantial agreement, and (2) it is more difficult for assessors to agree on content diversity than visual diversity of the SERP.

6 USERS' SATISFACTION WITH DIFFERENT LEVELS OF DIVERSITY

In this section, we aim to address **RQ2** (How search result diversity affect user satisfaction in image search with different search intents?). We conduct a user study to collect the satisfaction data and investigate how satisfaction scores change with the diversity levels and search intents.

6.1 Dataset

Based on the search tasks and SERPs we set in Section 4, we design a user study (shown on the bottom of Figure 2) to collect user satisfaction data. Participants are required to perform two warm-up search tasks first to get familiar with the experiment system and then finish 20 formal tasks. For each task, they are first presented with a search query and a short search scenario description to avoid ambiguity. After reading the description, one of the SERPs (*non-diversified* or *diversified*) we generate in Section 4.2 for this task is presented, which means that each participant sees only one SERP for a task. They are instructed to browse the SERP as they normally do where the query is not allowed to change. They could scroll to move the page up and down, click an image to view a high-resolution version of the image, and download it in the preview page. Finally, they are instructed to provide the SERP-level satisfaction feedback with one of five levels (1, not satisfied; 2, slight satisfied; 3, fair satisfied; 4, substantial satisfied; 5, very satisfied) based on how satisfied they were with the search experience in accomplishing the search task. Then they would be guided to continue to the next search task. Note that we ensure that each participant sees *non-diversified* SERPs in 10 tasks and *diversified* SERPs in another 10 tasks. Every two participants finish satisfaction feedback on all the 40 SERPs.

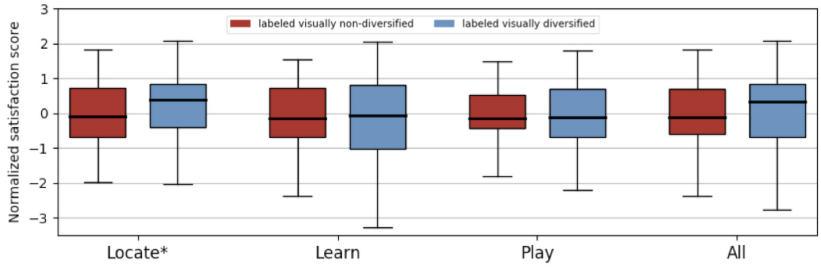
The satisfaction dataset involves 30 students (female=15, male=15) majoring in humanities and social science, engineering, and arts. All of them are reported as frequent users of Web search engines. It takes about 30 minutes to complete the user study, and we pay the participants about US$10 after they completed all the tasks seriously. We show the mean satisfaction scores distribution of 30 participants in Table 10. The mean satisfaction scores range from 3 to 4.35, which indicates satisfaction judgment may be quite subjective and different users may have different opinions. We regularize the satisfaction scores labeled by each participant into Z-scores according to Reference [8] as follows:

$$Z\text{-}score_i = \frac{Sat_i - Avg(Sat)}{Std(sat)},$$

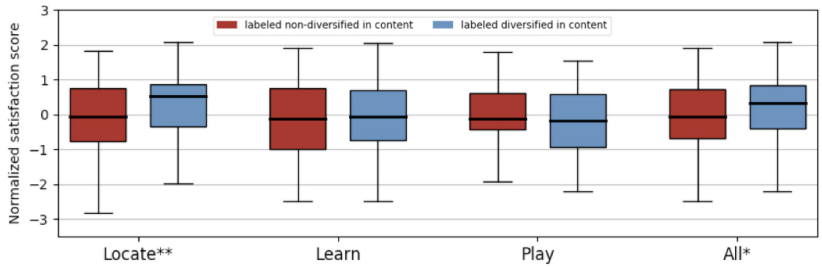
where Sat_i is one particular satisfaction score given by one user. $Avg(Sat)$ is the average of all satisfaction scores he/she labeled and $Std(sat)$ refers to the standard deviation. We use the normalization $Z\text{-}score_i$ as final user satisfaction score.

6.2 Satisfaction with Users' Perception of Diversity

With the user satisfaction dataset, we analyze how satisfaction changes according to users' perception of visual and content diversity and search intents.



(a)



(b)

Fig. 9. Distribution of user satisfaction scores for different tasks according to the user's perception of (a) visual diversity and (b) content diversity of the image SERPs.

Table 11. The Result of Significance Testing (p-value) on Satisfaction Scores with User's Perception of Visual and Content Diversity across Different Search Intents

Intent	p-value of t-test	
	Visual diversity	Content diversity
Locate	**0.035**	**0.002**
Learn	0.901	0.831
Play	0.842	0.260
All	0.091	**0.045**

6.2.1 Satisfaction with Visual Diversity. First, we look into how (annotated) perceived visual diversity affects user satisfaction. We draw the satisfaction score distribution as shown in Figure 9(a). The histograms from left to right are, respectively, satisfaction under *Locate*, *Learn*, *Play*, and all tasks. We find that users' satisfaction varies across non-diversified and diversified results and the difference varies across search intents. For *Locate* tasks, diversified result lists lead to higher satisfaction levels significantly (the p-value of t-test is shown in Table 11). However, for *Learn* tasks, the preference of diversified result lists is not significant. For *Play* tasks, users' preference of satisfaction does not have an obvious difference as well. As to all the tasks, diversified results lead to higher satisfaction evaluation in general but this is not significant (p-value = 0.091).

When users have a goal and want to download some images for further use, their information needs are specific. If the top results contain many similar images that they do not need, then users have to go deeper down the list of results to discover diverse views of the query. The longer users spend time on it, the more likely that they are not satisfied with the results. Therefore, for *Locate*

Table 12. Linear Regression: Coefficient of Variables

Features/Models	M1	M2	M3	M4	M5
precision (relevance)	0.0926	—	—	0.0910	0.0935
visual diversity	—	0.0332	—	0.0285	—
content diversity	—	—	0.0313	—	0.0338
MSE	0.2417	0.2489	0.2489	0.2410	0.2406

All coefficients are significant ($p < 0.01$); "M1–M5" refer to models constructed using different feature combinations. "—" indicates that this feature is not used in the current model. In terms of MSE (mean square error), all the models outperforms a random baseline significantly ($p < 0.01$).

intents, we should return diversified results to make sure the coverage of users' specific information needs. For *Learn* and *Play* tasks, diversity seems to have little effect on users' satisfaction.

6.2.2 Satisfaction with Content Diversity. Now we look into how user's perception of (annotated) content diversity affects satisfaction. The results are shown in Figure 9(b) and Table 11. Similarly to the result of visual diversity, for *Locate* tasks, diversified result lists lead to higher satisfaction levels significantly. For *Learn* tasks, the preference of diversified result lists is not significant. For *Play* tasks, non-diversified results tend to lead a little higher satisfaction level. As to all the tasks, diversified results lead to higher satisfaction evaluation but this is only significant with over 95% confidence (p-value = 0.045). The difference on *Play* tasks may associate with the relatively poor agreement of visual and content diversity for this task. In general, there is no significant preference for diversity for this type of exploratory tasks.

6.3 Does Diversity Indeed Affect Satisfaction?

A set of compounding factors can influence user satisfaction. For example, it has been shown in prior research that relevance is a significant factor of users' satisfaction [17]. Previously, we have demonstrated that users' satisfaction varies between non-diversified and diversified SERPs, especially for *Locate* search tasks. In this part, we use linear regression analysis to further examine the relationship between satisfaction and other factors, such as diversity (we mainly focus on) and relevance. We exploit our collected diversity and relevance annotations for this analysis. Since we mainly focus on diversity in this work, we use precision based on relevance, which is calculated as the average relevance measure with a 2-point relevance annotations, as our approach to quantify relevance. This has been validated to be one of the effective relevance-based offline evaluation metrics for image search from previous work Zhang et al. [57]. We report the mean squared error (MSE) and coefficient of the linear model in different feature combinations with a fivefold cross-validation in Table 12. There are 600 data points in our dataset, which are randomly divided into five groups. Coefficient stands for the magnitude of change caused by a one-unit change in the feature while other features being equal. It indicates how changes in relevance and diversity affect users' search experience. Note that all the features are normalized so that the coefficients are comparable.

We can observe in Table 12 that the precision score based on relevance has a high positive correlation with user satisfaction. This shows that to enhance user satisfaction, a search system should present relevant results. Besides relevance, users' perception of visual and content diversity also have a significant positive correlation with satisfaction. In terms of MSE, all the models perform significantly better than a random baseline approach. This demonstrates that diversity (M2 and M3) also relates to satisfaction in image search. We can also observe that when adding diversity to the model (M4 and M5), the MSE of the model decreases slightly (although the MSE differences are not significant). Overall, our results indicate that both relevance and diversity can be helpful in

predicting user satisfaction. This does not only reaffirm previous work on the contribution of relevance to user satisfaction but also establishes the relationship between diversity and satisfaction for the first time to our knowledge.

7 DISCUSSIONS AND LIMITATIONS

In this study, we explored the factors that affect users' diversity perception in image search and divided these prime factors into three categories: "(near) duplicate," content, and visual diversity. Based on the intent taxonomy of image search, we also examined how the perceived visual and content diversity of SERPs can reflect search satisfaction under different search intents. By collecting diversity annotations and a laboratory user study, we aimed to conduct a thorough investigation of users' perceived diversity and uncover the relationship between diversity and satisfaction in the context of image search.

First, to answer our first research question (**RQ1**), we generated candidate SERPs that vary in diversity with a heuristic algorithm. Through an interview after the diversity annotation, we identified the factor categories and their importance for diversity perception through an open-coded discussion method. Whether there exist (near) duplicated image results and the distance, presentation position of (near) duplicated images are the dominating factors that affect users' diversity perception. It is more noticeable when two duplicate images are ranked at the top positions of the SERPs. Visual presentation and content are another two prime aspects besides (near) duplicate in affecting users' perceived SERP diversity. The visual presentation may include the color distribution, hue, and background color of image results. The content aspect includes scene, composition, viewing angles, shots, as well as some query-dependent factors such as clothes, posture, and entity style. Since users may have very different views on diversity under query-dependent factors, it is found to be more difficult for assessors to agree on content diversity than visual diversity of the SERP.

Our second research question (**RQ2**) helped us understand how users' perceived visual and content diversity affect search satisfaction under different search intents. Based on the taxonomy (i.e., *Locate*, *Learn*, *Play*) of image search intent and the corresponding candidate SERPs we generated, we conducted a user study to collect the explicit user satisfaction data through interacting with those SERPs. Based on comparing satisfaction data with annotations of visual and content diversity, we found that users' satisfaction varied across labeled non-diversified and diversified SERPs. Search result pages with higher content diversity lead to higher user satisfaction levels significantly. In general, users prefer result lists with high visual diversity although this trend is not found to be significant. When further considering satisfaction across different search intents, search result pages with higher visual or content diversity lead to higher satisfaction levels significantly for *Locate tasks*. However, diversity seems to have very little effect on satisfaction for *Learn* and *Play* tasks.

Based on the experimental results, we conclude that (i) users' perception of diversity in image search can be affected by near-duplicate images, visual, and content-based features and (ii) a more diversified search result page can result in higher user satisfaction for certain types of tasks (i.e., *Locate*). We further discuss some of the rationales and limitations behind our experimental design and experimental findings.

7.1 Experiment Design

7.1.1 Task Design. Our search task design followed the image search taxonomy of *Locate*, *Learn*, and *Play* and this could have helped us make comparison across different search intents. However, those findings are limited to the small number of 20 search tasks designed by us, which may not reflect users' natural behavior. For example, one participant explained his frustrations due to the lack of knowledge of the information need, after finishing the search task "G-Dragon's girlfriend"

(Imagine you hear about the news that the Korean star "G-Dragon" has a girlfriend now and you want to have a look at her photos to kill time.): "I do not know who is G-Dragon, not to mention his girlfriend. I can not distinguish between relevant and irrelevant image results in this query. So I had to search for some other information before finishing this search task." Although given the user feedback, this is not the case for most of the search tasks, the artificially designed search tasks can potentially affect user behavior. To make our findings generalizable, it would be interesting to reaffirm our results in a more natural setting by conducting a large-scale field experiment to directly obtain data from users' own daily information needs.

Our search tasks are selected according to an existing image search intent taxonomy of *Locate*, *Learn*, and *Play* in this work. This taxonomy captures whether users have clear objects to find before they search and how image results satisfy their information needs (i.e., download for further usage or just browse the images). We find that users significantly prefer diverse SERPs only for *Locate* tasks. As users want to download some images for further use and their information needs are specific, it is understandable that diversified results are more likely to cover the specific information needs. These results suggest that when dealing with different tasks, the search engine should decide the correct scenarios to return a more diversified result list. Although the image search intent taxonomy we adopt helps us gain a more in-depth understanding on how satisfaction varies according to different intents, this intent taxonomy is very preliminary and cannot capture some fine-grained search scenarios, such as whether users care about the size of retrieved images, which may affect users' perception of search results. Another limitation of our study is the relatively small scale of our dataset, i.e., there are only 20 tasks in this work and these tasks do not capture such fine-grained information needs. We aim to collect more large-scale data for more tasks and investigate more fine-grained image search intent taxonomy in future work. For example, other search task taxonomies such as *Amorphous*/*Specific* and *Work*/*Entertainment* may also result in different diversity preference, which would be interesting to further investigate.

7.1.2 SERP Generation. In response to the manipulations of the diversity of candidate SERPs, we aim to control the SERPs to be as natural as possible for users to interact with. However, due to the nature of the controlled experiments, we also need to make compromises so that we can eliminate the effect of variables of no interest to us. Although this makes our findings more reliable or robust to certain biases, the findings are confined to the experimental settings. For example, we only reserve the top five rows of image results for one SERP, and query reformulation is not allowed during the user study that we collect explicit user satisfaction feedback. Although most of the users might browse only results from top five rows, users may, however, interact with more than five rows of images for exploratory search tasks, which may bring the experiment far from the real search scenario. Another potential concern of our study is the risk that our participants noticed the manipulation of search results in the satisfaction study even we try our best to avoid this effect. One participant said at the end of the study "I noticed that there are duplicate image results under some tasks, while the results of others are at high diversity levels. It seems not to be the original SERPs from the search engine. I wonder if you manipulate the diversity of SERPs." Despite this, we believe most of our participants did not notice the manipulations given the feedback. We would like to conduct a field study in our future work to make our experimental results and conclusions more reliable and universal, although it is more difficult to control the conditions of the experiments within this setting.

The candidate SERPs that we used during all the diversity annotation and user satisfaction experiments in this work were generated with simple SIFT (visual) feature and a heuristic algorithm. Although empirically we found this heuristic algorithm is effective and correlates very well with assessed diversity (see Table 7), there can be a potential bias in this approach. For example, due to

the nature of the SIFT algorithm, the SIFT feature we use in the heuristic algorithm can be biased to only the visual components of the images. This may render the SERPs to be more diversified from the visual perspective but not necessarily on the content perspective. This might be another explanation for the scenario that we found it is more difficult for assessors to agree on content diversity than visual diversity of the SERPs. We also found that although generally performing well, SIFT may sometimes underperform for calculating the similarity between certain images. For example, by manual inspections, we found that for some tasks (e.g., slides background), SIFT does not perform well in comparing the image results with no main objects. It would be interesting to import more image features such as color distribution, texture characteristic, and feature vectors extracted by the neural network for the more refined similarity calculation. Despite this, since the focus in this work is not to propose a novel and automatic method to manipulate the diversity of SERPs, we deem this heuristic algorithm sufficient for our purposes.

For future work, combined with the influencing factors of diversity perception we found in this work, we would like to formulate a formalized algorithm to generate SERPs that vary in diversity automatically. Similarly, the second research question relies strongly on the manual visual and content diversity annotations, which are affected by annotators' own points of views and are limited to small scale. It would be better to formulate an automatic and reliable algorithm for more refined diversity score calculation of the SERPs and thus enable the experiments to be large scale.

7.2 Diversity and Relevance of Search Results

When we generate diversified SERPs in Section 4.2, an interesting observation is that topical relevance and diversity can be conflicting with each other while both can significantly contribute to user satisfaction. During our SERP manipulations, some top-ranked images that are relevant to the query are eliminated, because they are too similar to a former image when we generated candidate SERPs. We find that the performance on relevance-based metrics of *non-diversified* and *diversified* SERPs are almost the same except for only the 5th (query: abdominal muscles) and 13th (query: Hebei province) search tasks. This may be caused by the difficulty of these two search tasks while there are not many relevant images. Despite this, it is interesting to observe that in Section 6.3, the experimental results reveal that indeed both relevance and diversity can contribute to user satisfaction. With a large-scale investigation in the future work, we would like to balance relevance and diversity and reveal a more delicate relationship between both diversity and relevance and how they interact with each other for user satisfaction in image search.

7.2.1 Relevance Loss. After the diversification process, the relevance of SERPs of certain queries might degrade. Therefore, according to whether there is loss on the performance of relevance (quantified by CG of row-average, as shown in Figure 3) after diversification, we categorize the queries into two groups: queries with relevance loss and queries without relevance loss. For those two query types, we, respectively, plot the distributions of normalized satisfaction scores for *non-diversified* SERPs and *diversified* SERPs in Figure 10. We find that when the relevance of *non-diversified* and *diversified* SERPs is almost the same (i.e., marginal or no relevance loss after diversification), *diversified* SERPs lead to higher satisfaction levels. On the contrary, when there is obvious relevance loss after the diversifying process, *Non-diversified* SERPs lead to higher satisfaction levels. For example, the normalized satisfaction scores on the *diversified* SERP for the fifth query decline by 0.41 (query: abdominal muscles, CG loss: 1.64). However, the normalized satisfaction scores on the *diversified* SERP for the eighth query increase by 0.27 (query: titanic, CG loss: 0.60). Given those two anecdotal examples, it demonstrates that users are able to tolerate some relevance loss when they want to get diversified search results. It would be interesting to

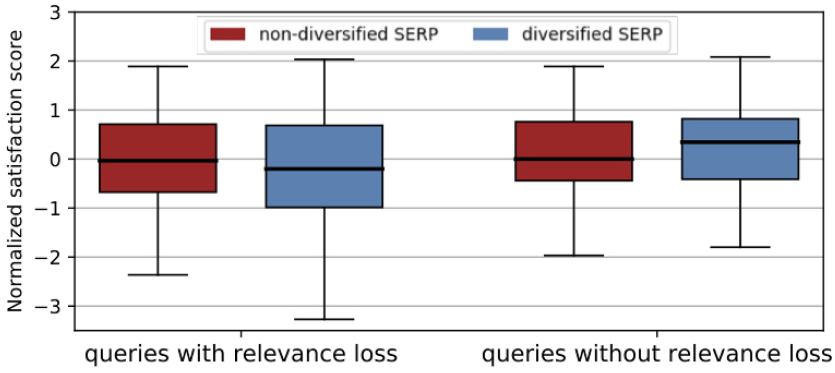


Fig. 10. Distribution of normalized satisfaction scores according to whether there is relevance (quantified by CG of row-average as shown in Figure 3) loss after the diversification process.

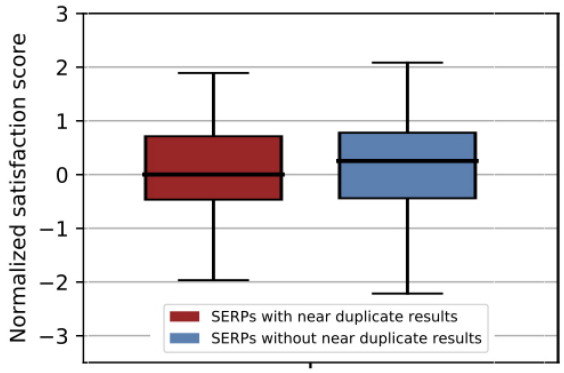


Fig. 11. Distribution of normalized satisfaction scores according to whether there exists near-duplicate image results on the SERPs. We only analyze 15 tasks, for which there are one SERP with near-duplicate image results (mostly one near duplicate) and one SERP without any near-duplicate image results.

investigate in large scale how much relevance loss users may be able to tolerate to get certain level of diversity in the future work.

7.2.2 Duplicate Results. Near-duplicate images are found to be the dominating factor that affect users' diversity perception greatly at first impression. This happens partly due to the fact that some of the candidate SERPs (i.e., non-diverse SERPs) are the original search result pages returned by the search engine. There exist near-duplicate image results that may be duplicate in visual appearance but maintain very different image textual tags or sizes or are originated from different websites (see Figure 1 for an example). This may suggest that for the image search engine to eliminate those visually near-duplicate image results, it may need to employ more complex strategies in trading off between textual diversity and visual diversity (near duplicate). We plot the distribution of the normalized satisfaction scores according to whether there exists near-duplicate image results on the SERPs in Figure 11. We only analyze 15 tasks, for which there is one SERP with near-duplicate image results (most of those only contain one near duplicate) and one SERP without any near-duplicate image results. We find that when there are no near-duplicate results on the SERP, the satisfaction scores are slightly higher (but not statistically significant). It indicates that although users focus on the near-duplicate results when assessing the diversity of

the SERPs, when interacting with image search results during the search process, they can accept some near-duplicate results and this does not affect so much on their satisfaction.

7.2.3 Visual and Content Diversity. With respect to the visual diversity judgment of SERPs, in most cases, the image results themselves are colorful and it is relatively easy to agree on the diversity level (fair agreement between assessors) based on those visual cues. However, for content diversity judgment, assessors' prior self-knowledge plays a decisive role, which makes it more difficult for assessors to agree on content diversity than visual diversity. This implies that to optimally diversify on image search results, the search engine should consider both the query-independent visual similarity and the query-dependent (and potentially personalized) aspects/subtopics that users may be interested in. Although both visual and content diversity can be correlated, it might be challenging for an algorithm to explicitly incorporate and optimally balance them.

In conclusion, we believe that we provide a solid foundation on revealing factors that affect user perceived diversity in image search and shed light on the relationship between diversity and user satisfaction for the first time. It is worth making more fine-grained investigations with a large-scale dataset in the future work.

8 CONCLUSIONS AND FUTURE WORK

Search diversification plays an important role in improving the quality of retrieval results. However, what factors affect users' perception of diversity and whether users prefer diversified result lists in image search have not been investigated. In this article, we conduct a thorough investigation into users' perception of diversity and the relationship between diversity and satisfaction.

We find that whether there exist (near) duplicated image results has the most significant impact on users' diversity perception, followed by the similarity in content and visual presentations. Users potentially have different criteria in defining visual and content diversity and therefore perceive differently on whether the SERP is diverse. We also notice that users' preference for diversity varies across different search scenarios in image search. While they want to collect information or save images for further usage (the *Locate* type queries), diversified result lists lead to higher satisfaction levels. Therefore, we should return diversified results to make sure the coverage of users' specific information needs. For *Learn* and *Play* tasks, diversity seems to have little effect on users' satisfaction. Users do not need to download images for further use in these tasks, and they even do not have a specific goal in mind. This leads to an uncertainty of diversity preference.

Our study is the first step in the necessity of diversified ranking in image search and may help commercial search engines decide the correct scenarios for a more diversified result list. Interesting directions for future work include proposing a novel algorithm to generate SERPs with different levels of diversity and automatically estimate the diversity levels of a SERP. Moreover, studying an automatic classification of search scenarios is worthwhile, which can be quite valuable for the search engines to decide whether to provide a more diversified result list in a specific search scenario.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. ACM, New York, NY, 5–14. DOI: https://doi.org/10.1145/1498759.1498766

[2] Paul André, Edward Cutrell, Desney S. Tan, and Greg Smith. 2009. Designing novel image search interfaces by understanding unique characteristics and usage. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, 340–353.

[3] Tevfik Aytekin and Mahmut Özge Karakaya. 2014. Clustering-based diversity improvement in top-N recommendation. *J. Intell. Inf. Syst.* 42, 1 (01 Feb. 2014), 1–18.

[4] Andrei Broder. 2002. A taxonomy of web search. *SIGIR Forum* 36, 2 (Sep. 2002), 3–10. DOI: https://doi.org/10.1145/792550.792552

[5] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*. ACM, New York, NY, 335–336. DOI: https://doi.org/10.1145/290941.291025

[6] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.* 11 (Mar. 2010), 1109–1135. http://dl.acm.org/citation.cfm?id=1756006.1756042

[7] Harr Chen and David R. Karger. 2006. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY, 429–436. DOI: https://doi.org/10.1145/1148170.1148245

[8] Ye Chen, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Meta-evaluation of online and offline web search evaluation metrics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. ACM, New York, NY, 15–24. DOI: https://doi.org/10.1145/3077136.3080804

[9] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 886–893.

[10] Atish Das Sarma, Sreenivas Gollapudi, and Samuel Ieong. 2008. Bypass rates: Reducing query abandonment using negative inferences. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*. ACM, New York, NY, 177–185. DOI: https://doi.org/10.1145/1401890.1401916

[11] Thomas Deselaers, Tobias Gass, Philippe Dreuw, and Hermann Ney. 2009. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'09)*. ACM, New York, NY, Article 39, 8 pages. DOI: https://doi.org/10.1145/1646396.1646443

[12] J. L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 5 (1971), 378–382.

[13] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.* 23, 2 (Apr. 2005), 147–168. DOI: https://doi.org/10.1145/1059981.1059982

[14] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. 2010. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the IEEE International Conference on Computer Vision*. 309–316.

[15] Qi Guo, Dmitry Lagun, and Eugene Agichtein. 2012. Predicting web search success with fine-grained interaction data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 2050–2054. DOI: https://doi.org/10.1145/2396761.2398570

[16] Jonathon S. Hare and Paul H. Lewis. 2013. Explicit diversification of image search. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval (ICMR'13)*. ACM, New York, NY, 295–296. DOI: https://doi.org/10.1145/2461466.2461513

[17] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: User behavior as a predictor of a successful search. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. ACM, New York, NY, 221–230. DOI: https://doi.org/10.1145/1718487.1718515

[18] Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (CIKM'13)*. ACM, New York, NY, 2019–2028. DOI: https://doi.org/10.1145/2505515.2505682

[19] Bernard J. Jansen. 2008. Searching for digital images on the web. *J. Doc.* 64, 1 (2008), 81–101.

[20] Bernard J. Jansen, Amanda Spink, Judy Bateman, and Tefko Saracevic. 1998. Real life information retrieval: A study of user queries on the web. *SIGIR Forum* 32, 1 (Apr. 1998), 5–17. DOI: https://doi.org/10.1145/281250.281253

[21] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retriev.* 3, 1–2 (2009), 1–224. DOI: https://doi.org/10.1561/1500000012

[22] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Comparing client and server dwell time estimates for click-level satisfaction prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. ACM, New York, NY, 895–898. DOI: https://doi.org/10.1145/2600428.2609468

[23] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. ACM, New York, NY, 113–122. DOI: https://doi.org/10.1145/2600428.2609631

[24] J. Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[25] Monica Lestari Paramita, Mark Sanderson, and Paul Clough. 2010. Diversity in photo retrieval: Overview of the ImageCLEFPhoto task 2009. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, Carol

Peters, Barbara Caputo, Julio Gonzalo, Gareth J. F. Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika (Eds.). Springer, 45–59.

[26] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. 2015. Different users, different opinions: Predicting search satisfaction with mouse movement information. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15)*. ACM, New York, NY, 493–502. DOI: https://doi.org/10.1145/2766462.2767721

[27] David G. Lowe. 2004. *Distinctive Image Features from Scale-Invariant Keypoints*. Kluwer Academic Publishers. 91–110 pages.

[28] Mathias Lux, Christoph Kofler, and Oge Marques. 2010. A classification scheme for user intentions in image search. In *Proceedings of the CHI'10 Extended Abstracts on Human Factors in Computing Systems (CHI EA'10)*. ACM, New York, NY, 3913–3918. DOI: https://doi.org/10.1145/1753846.1754078

[29] Rishabh Mehrotra, Imed Zitouni, Ahmed Hassan Awadallah, Ahmed El Kholy, and Madian Khabsa. 2017. User interaction sequences for search satisfaction prediction. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. ACM, New York, NY, 165–174. DOI: https://doi.org/10.1145/3077136.3080833

[30] Neil O'Hare, Paloma de Juan, Rossano Schifanella, Yunlong He, Dawei Yin, and Yi Chang. 2016. Leveraging user interaction signals for web image search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16)*. ACM, New York, NY, 559–568. DOI: https://doi.org/10.1145/2911451.2911532

[31] Richard L. Oliver. 1997. Satisfaction: A behavioral perspective on the consumer. *Asia Pac. J. Manage.* 2, 2 (1997), 285–286.

[32] Jaimie Y. Park, Neil O'Hare, Rossano Schifanella, Alejandro Jaimes, and Chin-Wan Chung. 2015. A large-scale study of user image search behavior on the web. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. ACM, New York, NY, 985–994. DOI: https://doi.org/10.1145/2702123.2702527

[33] Hsiao-Tieh Pu. 2005. A comparative analysis of web image and textual queries. *Online Inf. Rev.* 29, 5 (2005), 457–467.

[34] X. Qian, D. Lu, Y. Wang, L. Zhu, Y. Y. Tang, and M. Wang. 2017. Image re-ranking based on topic diversity. *IEEE Trans. Image Process.* PP, 99 (2017), 1–1.

[35] Filip Radlinski, Paul N. Bennett, Ben Carterette, and Thorsten Joachims. 2009. Redundancy, diversity and interdependent document relevance. *SIGIR Forum* 43, 2 (Dec. 2009), 46–52. DOI: https://doi.org/10.1145/1670564.1670572

[36] Filip Radlinski and Susan Dumais. 2006. Improving personalized web search using result diversification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, New York, NY, 691–692. DOI: https://doi.org/10.1145/1148170.1148320

[37] Vidyadhar Rao, Prateek Jain, and C. V. Jawahar. 2016. Diverse yet efficient retrieval using locality sensitive hashing. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR'16)*. ACM, New York, NY, 189–196. DOI: https://doi.org/10.1145/2911996.2911998

[38] D. M. Russell, D. Tang, M. Kellar, and R. Jeffries. 2009. Task behaviors during web search: The difficulty of assigning labels. In *Proceedings of the Hawaii International Conference on System Sciences*. 1–5.

[39] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY, 881–890. DOI: https://doi.org/10.1145/1772690.1772780

[40] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2010. Selectively diversifying web search results. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, New York, NY, 1179–1188. DOI: https://doi.org/10.1145/1871437.1871586

[41] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2011. Intent-aware search result diversification. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11)*. ACM, New York, NY, 595–604. DOI: https://doi.org/10.1145/2009916.2009997

[42] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a very large web search engine query log. *SIGIR Forum* 33, 1 (Sep. 1999), 6–12. DOI: https://doi.org/10.1145/331403.331405

[43] C. R. Snyder and S. J. Lopez. 2009. The Oxford Handbook of Positive Psychology. *Oxford library of psychology*.

[44] 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 (ICMR'15)*. ACM, New York, NY, 323–330. DOI: https://doi.org/10.1145/2671188.2749334

[45] Louise T. Su. 1992. Evaluation measures for interactive information retrieval. *Inf. Process. Manage.* 28, 4 (1992), 503–516.

[46] Benjamin W. Tatler and Benjamin T. Vincent. 2009. The prominence of behavioural biases in eye guidance. *Vis. Cogn.* 17, 6–7 (2009), 1029–1054.

[47] Hanghang Tong, Jingrui He, Zhen Wen, Ravi Konuru, and Ching-Yung Lin. 2011. Diversified ranking on large graphs: An optimization viewpoint. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*. ACM, New York, NY, 1028–1036. DOI: https://doi.org/10.1145/2020408.2020573

[48] Geoffrey Underwood and Tom Foulsham. 2006. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quart. J. Exp. Psychol.* 59, 11 (2006), 1931–1949.

[49] 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 (WWW'09)*. ACM, New York, NY, 341–350. DOI: https://doi.org/10.1145/1526709.1526756

[50] Bin Wang, Zhiwei Li, Mingjing Li, and Wei Ying Ma. 2006. Large-scale duplicate detection for web image search. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'06)*. 353–356.

[51] Gang Wang, Derek Hoiem, and David Forsyth. 2010. Learning image similarity from flickr groups using stochastic intersection kernel MAchines. In *Proceedings of the IEEE International Conference on Computer Vision*. 428–435.

[52] Hongning Wang, Yang Song, Ming-Wei Chang, Xiaodong He, Ahmed Hassan, and Ryen W. White. 2014. Modeling action-level satisfaction for search task satisfaction prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. ACM, New York, NY, 123–132. DOI: https://doi.org/10.1145/2600428.2609607

[53] Zhijing Wu, Xiaohui Xie, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. A study of user image search behavior based on log analysis. In *Proceedings of the China Conference on Information Retrieval*. Springer, 69–80.

[54] 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 11th ACM International Conference on Web Search and Data Mining (WSDM'18)*. ACM, New York, NY, 655–663. DOI: https://doi.org/10.1145/3159652.3159686

[55] Xiaohui Xie, Yiqun Liu, Xiaochuan Wang, Meng Wang, Zhijing Wu, Yingying Wu, Min Zhang, and Shaoping Ma. 2017. Investigating examination behavior of image search users. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'17)*. ACM, New York, NY, 275–284. DOI: https://doi.org/10.1145/3077136.3080799

[56] Yan Yan, Gaowen Liu, Sen Wang, Jian Zhang, and Kai Zheng. 2014. Graph-based clustering and ranking for diversified image search. *Multimedia Syst.* 23, 1 (2014), 41–52.

[57] 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 *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*. ACM, New York, NY, 615–624. DOI: https://doi.org/10.1145/3209978.3210059

[58] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. 2011. Image retrieval with geometry-preserving visual phrases. In *Computer Vision and Pattern Recognition*. 809–816.

Received August 2018; revised March 2019; accepted March 2019