A Large-Scale Study of User Image Search Behavior on the Web

Jaimie Y. Park*
KAIST, Korea
jaimie@kaist.ac.kr

Neil O'Hare
Yahoo Labs, Barcelona
nohare@yahoo-inc.com

Rossano Schifanella
University of Torino, Italy
rossano.schifanella@di.unito.it

Alejandro Jaimes
Yahoo Labs, New York
ajaimes@yahoo-inc.com

Chin-Wan Chung
KAIST, Korea
chungcw@kaist.edu

ABSTRACT

In this study, we analyze user image search behavior on a large-scale query log from Yahoo Image Search, based on the hypothesis that behavior is dependent on query type. We categorize queries using two orthogonal taxonomies (subject-based and facet-based) and identify important *query types* at the intersection of these taxonomies. We study user search behavior on a large-scale set of search sessions for each query type, examining characteristics of sessions, query reformulation patterns, click patterns, and page view patterns. We identify important behavioral differences across query types, in particular showing that some query types are more exploratory, while others correspond to focused search. We also supplement our study with a survey to link the behavioral differences to image search intent. Our findings shed light on the importance of considering query categories to better understand user behavior on image search platforms.

ACM Classification Keywords

H.1.2 User/Machine Systems: Human Factors; H.3.3 Information Systems: Information Search and Retrieval

Author Keywords

image search; search strategy; user behavior; query analysis; click analysis; search intent

INTRODUCTION

Understanding user behavior in web-scale image search is important because it provides valuable insights on content relevancy, opportunities for advertising, and for the design of the interaction and user interface. There is strong commercial interest by content providers in determining what types of images people are looking for, and a deeper understanding of how people search also benefits user interaction designers and system engineers when fine tuning the search engine.

*This work was carried out while Jaimie Y. Park was visiting Yahoo Labs Barcelona for a research internship.

CHI'15, April 18–23, 2015, Seoul, Korea.

http://dx.doi.org/10.1145/2702123.2702527

With today's high volume traffic on web search engines, one of the most common approaches to gain insights into user behavior is the analysis of query logs. Previous studies on web image search [1, 5, 8, 17, 20] focused on characterizing overall characteristics based on aggregated search logs. These findings help us to gain an overall picture of how users search for images on the web, but do not capture variation amongst different types of image queries; it is likely that behavior varies across query types, since user behavior in search is heavily dependent on task or goal [12]. Understanding such differences, and how they relate to user intent, is the main focus of this work. For instance, we posit that searches for 'Britney Spears' would, in general, exhibit different search patterns from those for 'desktop wallpaper', with the former likely to lead to casual browsing of celebrity images.

In this study, we identify what users search for on web image search, and study how user search behavior varies with query type, based on the logs from Yahoo Image Search (`images.search.yahoo.com`). To study the variation in behavior based on differences in query type, it is first necessary to have a clear understanding of the important query types in image search, which also allows us to identify the important dimensions of images that search engines need to understand. This motivates our first research question:

RQ 1. What do users search for on web image search?

Information seeking is a complex process, and to properly understand it, we also need to study *how* users interact with image search results. Previous work on image search has not examined how such behavior relates to query type; the insights that we can gain from answering this question can have important practical implications for search engines, since user interaction signals are used as implicit evidence for relevance feedback from users (e.g. relevant images get more clicks), and these insights can also enable adaptive results presentation based on query type. This motivates our second research question:

RQ 2. How does image search behavior differ depending on query type?

Human interactions with computers never take place in a vacuum, but always occur in response to some user need or goal, and image search is no exception [13]. As with any information service, to provide the best possible service it is crucial

to understand the underlying goals behind user behavior, motivating our final research question:

RQ 3. Can we associate query types with classes of search intent? How does search intent relate to behavior?

To summarize, the main contributions of this work are:

- We categorize a representative sample of image search queries using two complementary taxonomies, and identify a number of common image query types.
- We find a number of differences in search behavior based on query type; some query types are associated with exploratory, browsing-style behavior, while other query types users exhibit a more focused search.
- We emphasize three alternative sources of implicit feedback from search logs, *hovering over images*, *dwell time in the preview page*, and *click-through from image preview to referral website*. We show that position information is not always important for result clicks, and that click-through is heavily dependent on query type.
- We conduct a user survey to supplement our findings from the log-based analysis, and draw a connection between query type and known classes of image search intent.

In the next section we review related work, followed by a dataset description. We then characterize a sample of image search queries according to two orthogonal taxonomies, and identify important *query types* for further analysis. We then examine query type based differences in search behavior and present the results of a qualitative survey. Finally, we discuss the implications of our results and conclude the study.

RELATED WORK

'What' Images People Search for. Understanding what users search for allows us to identify the aspects of images that should be represented to support search, and several studies in the past have attempted to classify user queries on web image search engines. Pu [17] classified the 1,000 most frequent image queries based on a proprietary subject-based categorization scheme, and found that a majority of the queries were in the entertainment domain. Jansen [8] classified queries based on three existing non subject-based image query classification schema, focusing on whether users were searching for people, locations, etc., and on whether the search was about unique instances or non-unique instances, which is closely related to the facet-based Shatford-Panofsky categorization framework [18]. This framework was also used by Armitage and Enser [2] to classify image queries in offline multimedia archives.

The above studies, and other similar studies [6, 8, 17, 20], agree in their main findings that web image search is dominated by people queries or queries within the *Arts & Entertainment* category. They either explicitly focus on the most popular queries or do not reveal details of how they sample their queries, suggesting a possible bias towards popular queries. In this study we avoid such bias by taking a stratified sample from the entire query distribution; we further compare popular and rare queries to examine the differences between the two. The characterizations that emerge from subject-based or facet-based characterizations can be very broad (e.g. 'Arts & Entertainment'). To uncover more fine-grained categories that cover a significant amount of query traffic, we look at the query types that emerge at the *intersection* of subject-based and facet-based categorizations.

'How' People Search for Images. Following on from the work of Goodrum and Spink [5], researchers started working on characterizing image search behavior on web search engines [1, 5, 6, 8, 15, 17, 20]. These studies characterize the general behavior of users on image search platforms based on aggregated search log data, measuring features like session length, the number of result pages viewed, the number of results clicked per session and query reformulation patterns. They typically compare image/multimedia search behavior with general web (text) search behavior, finding that image searches lead to more clicks and deeper exploration of the search results (search depth), and conclude that image search tends to be more 'exploratory' and requires greater interactivity [1, 10]. Other work focus on search behavior on specialized image sharing platforms like Flickr [14].

Andre et al. [1] argue that, while image search tends to be more exploratory than text search, image searches can also be goal-directed. What factors affect this were not explored in the study, nor in any other study of web image search behavior. We posit that there should be significant differences in search behavior depending on the query type. Also, the above studies limit their behavior analysis to the statistics listed above (session length, number of result pages viewed, etc.), ignoring important interactions such as hovers on images, the relationship of clicks with rank position, and interactions with the image preview page (the page displayed after the user clicks on a result), all of which constitute a large portion of user interaction with image search engines. In our work, we address this gap by considering various user interactions with image search engines, and by conducting an in-depth study of behavioral differences based on query type.

'Why' People Search for Images. The general goals and tasks that motivate search provide the context for search behavior. Broder [3] proposed a taxonomy of *intent* for web search, which was adapted to image search by Lux et al. [13]. No previous work has attempted to link these classes of image search intent to specific query types, or to behavior observed in image search logs; in this work, we make a first attempt at linking intent, behavior and query type through a user survey that elicits the testimonials of real users.

DATASET

From the Yahoo Image Search server logs, we take as a sample all the queries issued in the U.S., across all device types (desktop, mobile, etc), during a continuous period in the fall of 2013[1]. We remove adult queries using an automatic classifier, and sessionize the data by partitioning a user's actions

[1]We do not reveal the exact time period, as revealing both volume and time period would disclose commercially sensitive information (i.e. volume of searches per day). However, the period under consideration does not involve major festivities, e.g., Christmas or Thanksgiving, that could introduce a problematic bias.

Number of Searches	102,534,341
Number of Unique Queries	36,103,126
Number of Sessions	34,715,204

Table 1: Dataset Description.

into separate *sessions* when the time between consecutive actions exceeds 30 minutes. The filtered, sessionized, sample contains approximately 102M searches, and 36M unique queries (Table 1). To group unique queries, we cast all queries to lower case and remove punctuation.

The query traffic distribution, as expected, has a long tail: 75% of unique queries (27M of 36M unique queries) were issued only once, and they account for approximately 25% of the traffic in the sample; the other 25% of queries account for 75% of all traffic. At the head of the distribution, the top 1% of unique queries (360K out of 36M) account for 46% of all traffic, and the top 10% account for 60% of all traffic.

From the logs, we extract data for events in two different page types: the *search results page* and the *image preview* page. The *image preview* page, a variation of which exists in the image search site of each of the major U.S. search engines, is an enlarged preview of an image, shown after the image is clicked on the search results page, with *next* and *previous* navigation buttons for further exploration of the results. A link to the referral website where the original image can be found is also displayed. We extract the following fields from each entry in the log: session id, timestamp, query string, anonymous user identifier, page type, and event type (i.e. pageview, click). For pageview events we extract the urls of all thumbnail result images displayed on the page, and the position they were displayed at. For click events, we have information about the type of click (e.g. click, hover), and the URL and rank position of the clicked image.

CATEGORIZATION OF IMAGE QUERIES

In this section we address *RQ 1*, *'What do users search for on web image search?'*. We begin by outlining our methodology, and then present our results in the section *'What Users Search for in Web Image Search'*.

Previous studies on what people search for either focus on only the most popular queries or do not give full details of how the queries were sampled, making it unclear if their findings are representative of all queries. Focusing only on popular queries would give a biased view (e.g. too much focus on celebrities), obfuscating the variety among less frequent queries. In this work, we take a sample that represents the entire distribution. Additionally, given that the long tail of rare queries takes up such a large portion of the traffic (e.g. singleton queries make up 25% of the traffic volume), we believe it is important that we understand what types of queries it consists of, and how this compares to the popular queries at the head of the distribution. This motivates our first subquestion:

RQ 1.1. Are there differences in the types of queries found in the head and the tail?

Previous studies conducted either subject-based classification or facet/aspect-based classification, leading to characterizations that are overly broad. For instance, a subject-based taxonomy may tell us that *Arts & Entertainment* queries are frequent, but this could refer to a wide range of instances, from a famous actor to a book title. By looking at the intersection of two taxonomy types, we can uncover finer-grained, but popular, query types. This motivates a second subquestion:

RQ 1.2. What are the frequent query types at the intersection of subject- and facet- based categorization?

Methodology for Query Categorization

For categorization, we take a sample of 1,000 queries to annotate manually. To reduce the sampling error that can arise from a random sample, we use a stratified sampling method. All search queries are partitioned into 1,000 equal-sized strata, where the n^{th} stratum represents the n^{th} permile[2] of the query traffic distribution[3]. In other words, each stratum is composed of queries that account for the same amount of traffic (i.e. the 1^{st} stratum represents queries covering the top 0.1% of the traffic, etc.). We sample a single query from each of the 1,000 strata. This sampling strategy ensures that the set of annotated queries evenly covers the entire traffic distribution, and that by averaging over the annotated queries we get an average that represents all query traffic. It also allows us to separately analyze different portions of the query distribution (e.g. the first 100 query samples represent the top 10% of the distribution). We categorize the queries based on two complementary classification schemas described below.

IPTC Subject Taxonomy. Since there is no standard subject-based taxonomy for image queries, we chose the IPTC subject code taxonomy[4], which is often used for classifying online content, and has previously been used for images [21]. The taxonomy includes 17 top level subject-based nodes (e.g. *Arts, Culture & Entertainment*, *Lifestyle & Leisure*). Through the process of manual annotation, we identified an additional root category, *Nature*, not covered by the IPTC taxonomy. Also, the most frequent root categories, *Arts, Culture & Entertainment* and *Lifestyle & Leisure*, cover a range of subtopics, so we use IPTC subcategories for these. We also add three additional subcategories of *Lifestyle & Leisure* that we identified as important: (*Clothing & Accessories*, *Automobiles*, and *Graphics/Clipart*). The taxonomy is shown in Table 3.

Shatford-Panofsky Framework. The Shatford-Panofsky approach for image indexing was introduced by Shatford [18], based on a previous theoretical work [16], and was later adopted for classifying queries in image archives [2]. It was found to be useful in characterizing the way people formulate queries and revealing their visual information needs. The approach, summarized in Table 2, characterizes images/queries based on four *facets* (*who, what, where when*) and three *aspects* (*specific, generic, abstract*). Note that the *who* facet represents any entity, not just people, while the *what* facet

[2]permile = 1000-quantile

[3]Note that stratification was performed on all the searches, not on unique queries, so that a stratum may contain redundant queries (e.g. the first head stratum was entirely filled with 'miley cyrus').

[4]http://cv.iptc.org/newscodes/subjectcode (accessed Jan 14th, 2015)

	Specific	Generic	Abstract
Who	individually named person/group/thing (e.g. 'miley cyrus')	kind of person/group/thing (e.g. 'tiger')	mythical or fictitious being (e.g. 'pikachu')
What	individually named event/action (e.g. 'presidential election 2016')	kind of event/action/condition (e.g. 'commencement ceremony')	emotion or abstraction (e.g. 'cold')
Where	individually named geographical location (e.g. 'seoul')	kind of place: geographical/architectural (e.g. 'city')	place symbolized (e.g. 'paradise')
When	linear time: date/period (e.g. 'april 18th, 2015')	cyclical time: season/time of day (e.g. 'summer')	emotion symbolized by time (e.g. 'dark age (of life)')

Table 2: Facet/Aspect Categorization Schema.

represents actions, conditions and events (i.e. *what* that entity is doing). A query could instantiate the *who* facet and the *specific* aspect (e.g. 'miley cyrus') or the *who* facet and the *generic* aspect (e.g. 'tiger'). We further subcategorize the queries in the *what* facet as an *action*, *event* or *condition*. Similarly, we subcategorize queries in the *who* facet using the top level of the Schema.org entity type taxonomy[5].

Manual Annotation Procedure. To ensure an accurate categorization of our sampled queries, we manually annotate them into the two taxonomies. Manual annotation is especially important for the Shatford-Panofsky schema, where automatic classification is difficult. Three of the authors of this paper manually annotated the queries using an iterative process. Initially, a small test sample of queries, independent of the 1,000 sample, were annotated by all three annotators. The results were compared and conflicts discussed: this process was iterated until a stable set of annotation guidelines was agreed upon. Finally, each of the annotators annotated a random subset of the 1,000 query sample into both taxonomies. Any query identified as 'difficult to annotate' was discussed by the 3 annotators until a consensus was reached.

What Users Search for in Web Image Search

IPTC Categorization Results. Table 3 shows the IPTC subject categorization results. Subject categories covering less than 3% of the query volume are omitted from the table. *Lifestyle & Leisure* and *Arts, Culture & Entertainment* are the predominant subjects, covering more than 70% of queries, followed by *Sport* and *Nature*. Within *Lifestyle & Leisure* the most popular sub-categories are *clothing & accessories* (e.g. 'custom tuxedos'), *graphics/clipart* (e.g. 'android funky icons'), *automobiles* (e.g. 'ford explorer') and *house & home* (e.g. 'kitchen cabinets').

Shatford-Panofsky Categorization Results. Table 4 shows the Shatford-Panofsky classification results, with searches for specific people or things (*specific-who*) by far the most frequent, covering 57% of searches. Within this, people or product searches (e.g. 'miley cyrus', 'iphone 5c') are the most common. *Generic-who* queries (e.g. 'flower', 'wedding rings') are also very frequent, covering 41% of searches. The *who* facet is clearly the most important facet in image search, covering almost 93% of all searches. The next most frequent facet is *what*, covering 21% of searches. Unlike the *who* facet, *what* queries are more likely to be generic; users, in particular, often use generic conditions (7.4%) as modifiers for who queries (e.g. 'blue quilt'). It is also notable that the *when* and *where* facets mostly appear together with other facets, as modifiers of the terms associated with *who* and *what*. For example, the query 'flowers to plant in september' refers to a *generic-who* entity (flowers) modified by a *generic-what-action* (to plant) and a *generic-when* description (in september). 87% of *specific-where* queries, and 95% of *specific-when* queries, are multifaceted.

Differences between Head and Tail. Tables 3 and 5 also allow us to answer *RQ 1.1*, *'Are there differences in the types of queries found in the head and the tail?'*, by comparing

Category	All %	Head %	Tail %
Lifestyle & Leisure	**36**	2	43
Clothing & Accessories	**8**	1	10
Graphics/Clipart	**5**	0	4
Automobiles	4	0	5
House & Home	4	0	6
Gastronomy	3	0	4
Tourism	3	0	4
Others	9	1	9
Arts, Culture & Entertainment	**35**	85	16
Cinema	**15**	52	5
Television	**14**	39	4
Music	**7**	17	3
Fashion	4	13	1
Animation	3	0	1
Others	5	8	5
Sport	**7**	6	7
Nature	**6**	4	4
Science & Technology	4	0	3
International Interest	3	0	0
Health	3	0	4

Table 3: IPTC Categorization.

Category	Specific %	Generic %	Abstract %	All %
Who	**57.1**	**41.1**	5.1	**92.6**
Person	**28.2**	2.4	2.0	
Product	**15.5**	**14.6**	0.4	
Place	2.7	1.1	0.3	
Organization	5.3	0.4	0.0	
Animals/Plants	0.3	5.3	1.2	
Others	3.9	16.5	1.0	
What	2.6	**13.4**	5.4	**21.0**
Action	0.4	3.3	0.1	
Event	2.1	2.6	0.0	
Condition	0.1	7.4	5.1	
Others	0.0	0.2	0.1	
When	3.7	1.3	0.0	5.0
Where	5.4	0.5	0.0	5.8
All	**60.7**	**45**	10.3	

Table 4: Shatford-Panofsky Categorization.
(As queries can be multi-faceted, rows and columns do not sum to 100%, or to the values for 'all'.)

[5]http://schema.org/docs/full.html (accessed Jan 14th, 2015)

Category	Specific %		Generic %		Abstract %		All %	
	(h)	(t)	(h)	(t)	(h)	(t)	(h)	(t)
Who	94	48	4	50	0	5	98	89.6
What	0	4	1	22	2	6	3	31.6
Where	0	12	0	1	0	0	1	6.4
When	0	5	1	6	0	0	0	12
All	94	54.4	5	56	2	11.2	-	-

Table 5: Shatford-Panofsky Categorization for Head (h) and Tail (t) Queries.

head and tail queries. We consider head queries as those in the top 10% of the distribution, and tail queries as those in the bottom 25% (we choose the bottom 25% because this is made up entirely of singleton queries). Significant differences emerge in the distribution of categories for head and tail queries. For IPTC subjects, the head is dominated by *Arts, Culture & Entertainment* queries (85%), followed by *Sport* and *Nature*. Tail queries cover a much more diverse set of topics, with *Lifestyle & Leisure* covering 43% of queries, followed by *Arts, Culture & Entertainment* covering 16%, and a non-negligible amount of queries related to *Sport*, *Nature*, and *Health*.

There are also differences in the Shatford-Panofsky categories between head and tail (Table 5), with the head dominated by *specific-who* queries (94%). In the tail this drops to 48%, with 50% covered by the *generic-who* facet. There are also more *what* and *where* queries in the tail, and a far greater proportion of *generic* queries.

These huge differences in the nature of head and tail queries emphasize the need to study the entire query distribution, and not just the popular queries.

Intersection of Subject and Facet Taxonomies. To answer *RQ 1.2*, *'What are the frequent query types at the intersection of subject- and facet- based categorization?'*, we look at the frequent query types at the intersection of the two taxonomies. We only consider the intersections of subjects and facets that cover at least 3% of the query traffic, and label each of these as a specific *query type* (e.g. *celebrity* queries have been annotated as cinema, television, movie or sports AND specific-who-person).

From Table 6 we see the following common *query types* at the intersection of the two schema, which between them account for 43.7% of traffic: *celebrities*, *graphics/clipart*, *fashion items*, *animals*, and *automobiles*. *Celebrity* and *automobile* queries both cover the *specific-who* facet, while the others cover *generic-who*.

A further inspection of the head and tail queries shows that *celebrities* queries, at the intersection of *Arts, Culture & Entertainment* or *Sport* and *specific-who*, are mainly drawn from the head of the distribution, while *graphics/clipart* and *fashion items* are more typical of the tail.

USER BEHAVIOR IN IMAGE SEARCH

In this section, we analyze user behavior in image search, focusing on behavioral differences between the query types identified in the previous section, addressing *RQ 2 'How does image search behavior differ depending on query type?'*.

Subject	Facet/Aspect	Query Type	%	Example
Cinema Television Music Sport	Specific/Who -Person	Celebrities	26.4	brad pitt kaley cuoco beyonce michael jordan
Graphics /Clipart	Generic/Who -Others	Graphics /Clipart	4.5	fall wallpaper, pumpkin clipart
Clothing & Accessories	Generic/Who -Product	Fashion Items	6.4	patent handbag, custom tuxedo
Nature	Generic/Who -Animals	Animals	3.4	dogs, wolf spiders
Automobiles	Specific/Who -Product	Automobiles	3.0	ford explorer, chevy malibu

Table 6: Popular Query Types Identified through Subject & Facet/Aspect Categorization.

Previous work has suggested that image search is more exploratory than web search [1]. Motivated by this, we are interested in understanding more fully to what extent this exploratory behavior applies to all query types:

RQ 2.1. For which query types can we find further evidence of exploratory behavior in image search?

Existing studies on web image search behavior focused on interactions with the search result page, dwell time on the results page, clicks on result images, and the number of result pages viewed. Other important interactions with search engines, such as image-hovers and interactions with the preview page, have not yet been explored. Given the large volume of such data (e.g. hovers are much more frequent than clicks), we also ask the following subquestion:

RQ 2.2. Do additional interaction features on modern search engines provide useful insights for understanding query type based behavior differences?

In the subsections below, we study session-level statistics, query reformulation patterns, and interactions with the preview page. The main research question *RQ 2* and the subquestion *RQ 2.1* are covered by all the subsections, whereas the subquestion *RQ 2.2* is covered in *Click/Hover-Through Rates*, *Click/Hover Entropy*, and *Comparing Clicks and Hovers*, which deal with result-interaction via hovering, and *Preview Page Interactions*, which deals with interactions with the image preview page.

To study query type based behavioral differences, we first examine the number of search pages viewed, query reformulation patterns, and number of result clicks, which were used in previous work to suggest that image search is more exploratory than web search [1]. In addition, we study features not explored in other studies, but which constitute a large portion of user interaction in image search, such as hover-through, the relationship between click-through and rank position, dwell time on preview page, and click-through from the preview page to the referral website.

For a large-scale analysis, we expand the set of queries associated with each *query type* identified at the intersection of subjects and facets. We run a named entity detector against the entire query sample to classify *celebrity* and *automobile* queries, and a text classifer for *animal* queries. We extract

Query Type	Searches	Unique Queries	Sessions	Searches/ Query
Celebrities	12,130,246	7,367	5,625,328	1646.5
Graphics/Clipart	1,487,864	643,299	574,299	2.3
Fashion Items	1,850,754	1,039,390	974,139	1.8
Animals	420,493	12,410	253,003	33.9
Automobiles	239,673	9,849	150,759	24.3

Table 7: Expanded Dataset by Query Type.

graphics/clipart queries as those containing a list of curated keywords[6] and *fashion item* queries based on the keywords from Google Product Taxonomy[7], under the *Apparel and Accessories* > *Clothing* category. Unlike our query characterization, which considered all device types, for this analysis we wish to to control for behavior differences based on device, so we only consider queries and sessions on desktop devices. The resulting dataset, which will be the testbed for our analysis in this section, is summarized in Table 7. As can be seen in the *Searches/Query* column, the query types in this dataset are a mixture of very popular queries (*celebrities*), moderately popular queries (*animals*, *automobiles*), and rare queries (*graphics/clipart* and *fashion items*).

Session Level Statistics

We start by looking at query type differences in session-based statistics, which give a high-level overview of search engine interactions.

Query Type	Session Duration (mins)	Queries per Session	Queries per Minute	Reformulation (%)	Search Depth (pages)
Celebrities	8.59	4.09	0.48	60.09	2.76
Graphics/Clipart	16.91	5.28	0.31	70.97	3.31
Fashion Items	15.29	4.99	0.33	61.44	2.96
Animals	13.55	4.88	0.36	60.31	2.32
Automobiles	10.87	3.78	0.35	54.04	2.66

Table 8: Session Statistics by Query Type.

From Table 8 we can see a clear difference in the duration of sessions containing *celebrity* queries, where the average length of a session is noticeably shorter than for all other query types. Table 8 also shows that *graphics/clipart* sessions have the largest number of queries and the longest average session length. Although longer sessions tend to contain more queries, queries are not issued at the same rate for each query type; for some query types, it takes longer for the same number of queries to be issued. *Celebrity* search sessions contain 0.48 queries per minute, which is the largest among all, suggesting that users spend less time interacting with the results, and move on to new/refined queries more quickly. Consistent with the greatest number of queries per session, there is a greatest tendency to reformulate queries for *graphics/clipart* sessions (70.97% of sessions contain more than one query) and least for *automobile* sessions (54.04%). Lastly, we compute search depth, based on how many pages of results a user explores, to see how deeply users explore the results: *graphics/clipart* and *fashion item* queries show the greatest search depth. These query types are associated with longer, deeper search sessions with more reformulations, suggesting they are relatively more focused, while other query types (*celebrities*, *animals*, *automobiles*) are characterized by shorter session lengths, less reformulations, and shallower search depths.

[6]the following keywords were used: 'graphic', 'clipart', 'clip art', 'wallpaper', 'background', 'icon', and 'illustration'

[7]www.google.com/basepages/producttype/taxonomy.en-US.xls (accessed Jan 14th, 2015)

Query Type	Adding %	Deleting %	Partial %	Complete %
Celebrities	15.4	0.6	25.1	58.8
Graphics/Clipart	18.3	3.8	47.0	30.9
Fashion Items	16.5	2.9	44.6	36.1
Animals	13.0	2.2	31.9	52.9
Automobiles	16.0	2.6	40.4	41.0

Table 9: Query Modification type by Query Type.

Query Reformulation

Previous studies show that query reformulation occurs frequently on image search platforms [1], and our data also provides evidence to support this, with approximately 60% of sessions involving more than one query. In this subsection, we look more deeply into the types of query reformulation used and how this relates to query type. Query reformulation is normally defined as any modification that a user makes to the initial query in hope of finding better results [7]. We adopt the approach of Jansen [9] and distinguish the following types of reformulation between consecutive queries: (i) *adding terms*, where one or more terms are added to the original query, (ii) *deleting terms*, where one or more terms are deleted from the original query (iii) *partial change*, where at least one term has been changed, but the updated query is neither a superset or a subset of the original, and (iv) *complete change*, where the two queries have no terms in common.

Table 9 shows the extent to which each type from reformulation occurs across different query types in our dataset. For all categories, many reformulations involve partial changes or complete changes to the preceding query. Term deletion is the least common type of reformulation, and adding terms is also relatively rare. We find noticeable differences in reformulation strategies for *celebrity* queries, which have a much higher proportion of complete changes, and lower proportion of deletions and partial changes, than other query types. *Graphics/clipart* queries have the most partial changes and adding terms, and the least complete changes. Table 10 shows some examples of the frequent query reformulations for each

Query Type	Popular Query Reformulations
Celebrities	*top:* elizabeth hurley → katherine webb charlie hunnam → dakota johnson christie brinkley → kate upton
Graphics/Clipart	*top:* school clip art → school clipart fall clipart → halloween clipart dallas cowboys → dallas cowboys wallpaper
Fashion Items	*top:* short haircuts → short hairstyles short haircuts → short haircuts 2013 short shorts → yoga shorts
Animals	*top:* eagle attacking deer → puppy doe cat → dog puppies → kittens
Automobiles	*top:* 2014 corvette → ferrari porsche 918 spyder → porsche carrera gt gold-plated lamborghini → kim zolciak

Table 10: Frequent Query Reformulations by Query Type. (The most frequent reformulation pair is denoted as *'top'*)

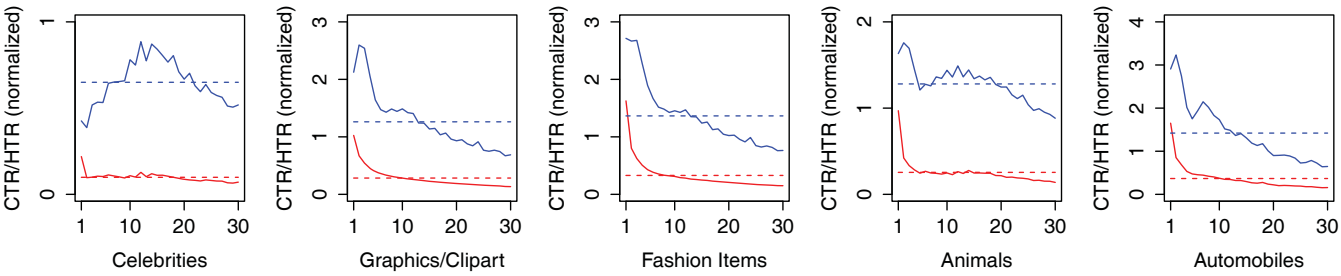

Figure 1: Click(bottom/red)/Hover(top/blue)-through Rate for Images for Top 30 Position Ranks by Query Type. (CTR/HTR are normalized to hide commercially sensitive data. Dotted lines denote the average over all positions.)

query type. The results suggest a more exploratory, browsing-like, behavior for *celebrity*, *automobile*, and *animal* queries, which have more complete changes to a different entity of the same type (e.g. 'puppies'→'kittens'). In contrast, for *fashion item* and *graphics/clipart* searches, users tend to refine queries through adding terms or partial changes, which serves as evidence of focused search behavior (e.g. 'fall clipart'→'halloween clipart').

Interaction with Search Results

In this section, we analyze how users interact with results after a query has been issued, based on explicit interactions with search results. Understanding this behavior is very important, as it can be an indication of user satisfaction as well as relevance of search results, and is heavily used by search engines as relevance feedback data. We focus on the following types of interaction with search results: *clicks* and *hovers*[8] on search result images, and interactions with the image preview page. We note that approximately 10% of our data consists of searches where no follow up action takes place whatsoever: the user inputs a query, views the first returned page, and ends the session with no further action.

Click/Hover-Through Rates. We now measure click-through rate (CTR), which has been shown to have a strong correlation with image relevance [4, 19], as well as hover-through rate (HTR). For each *(image,query,position)* triple, where position is the ranking of the image in the result list, we calculate CTR/HTR, the number of clicks/hovers on the image divided by the number of image views. In Figure 1, we plot the average CTR and HTR of images at the top 30 positions for each query type. HTR is significantly higher than CTR for all query types; indeed from the *click/hover ratio* in Table 11 we can see that it can be as much as 10 times higher for *celebrity* and *animal* queries. In terms of CTR (bottom/red lines), a sharp decay is observed over the top 5 images for all query types except *celebrities*, followed by a slower decay. For *celebrities*, the CTR stabilizes after the second result, and only starts to decay again after position 16. We also see a similar flat line for CTR on *animal* queries. For both *celebrity* and *animal* queries, the HTR is also somewhat unusual in that it increases with position, albeit after an initial drop in the case of animal searches. These CTR and HTR plots suggest that for *celebrity* searches, users want to browse a set of images; the relatively smaller impact of the ranking suggests that users are consuming the results (at least those in the top 30) as an unordered set, and not as a ranked list. The fact that CTR/HTR shows a gradual decrease for *graphics/clipart* and *fashion items* again suggests a more focused search, as position has more influence on CTR.

It is also very important to note that the overall CTR and HTR vary widely by query type. *Celebrity* searches, for example, have very low click-through in comparison with other query types; this suggests that these queries are often satisfied by the thumbnails on the results pages, without the need for further interaction, again suggesting a browsing-like behavior. These differences in CTR and HTR can also have very important implications for interpreting click data as an implicit relevance signal, as we will discuss later.

Click/Hover Entropy. To measure the amount of variation in the search results users click/hover on across query types, we calculate click/hover entropy, using the standard measure

$$Click/HoverEntropy(q) = \sum_{i,q} -p(i|q)log_2p(i|q),$$

where $p(i|q)$ is the probability of image i being clicked for query q. Since infrequent queries would have a low entropy and so generate bias, we compute entropy only for queries that were issued at least 20 times. Table 11 shows that hover entropy values are higher than click entropy values, which is to be expected since there are many more hover actions. The highest entropy is observed for *celebrity* and *animal* queries, which means it is harder to predict what image a user will click/hover on for these types. This again hints at the undirected, browsing-oriented nature of such queries in comparison to *graphics/clipart* and *fashion item* queries.

Query Type	Click Entropy	Hover Entropy	Click/Hover Correlation	Click/Hover ratio
Celebrities	4.75	6.47	0.49	0.09
Graphics/Clipart	4.03	6.07	0.44	0.13
Fashion Items	3.93	5.46	0.51	0.12
Animals	4.32	6.70	0.56	0.09
Automobiles	4.11	5.98	0.58	0.15

Table 11: Click/Hover Entropy, Correlation, and Ratio by Query Type.

[8]Hovers refer to a user placing the cursor on the thumbnail of a result image. Although this feature is not available on all image search engines, at the time of writing 2 out of the 3 top U.S. search engines (Bing and Yahoo) had implemented this user interface feature.

Comparing Clicks and Hovers. Given that there are many more hovers than clicks, and that clicks are highly correlated with relevance [4, 19], in this section we ask how similar hovers are to clicks, which will also tell us whether hovers are correlated with relevance. To measure the extent to which users hover over the same images that they click on for a given query, we first rank the images based on the number of times they were clicked or hovered on for a query, and compute the correlation between the click and hover ranks. We remove sparse data by only considering *(query,image)* pairs with at least 10 clicks, and only queries with at least 10 such images. Table 11 shows the correlation between clicks and hovers, which on average is moderate, at around 0.5. *Graphics/clipart* queries have the lowest correlation, consistent with our earlier interpretation that for this query type, users may be performing a more focused search, hovering over images for further inspection, without necessarily clicking on them. Finally, we calculate the click/hover ratio for each query type. The results in Table 11 show that, consistent with our other findings, *celebrity* and *animal* queries have the lowest proportion of clicks, compared with hovers, suggesting that hovers are more often used as a way to consume these images.

Preview Page Interactions. Modern search engines typically allow users to preview images after clicking on them, showing a maximized view of the image, along with next/previous buttons, on a new page which we refer to as the image preview page. This page also has a link to the referral website where the image can be viewed in its original context[9]. Table 12 shows that an average of 17 or more images are viewed on this page for every preview page visit (i.e. for every result click); this number is consistent across most query types, and the high volume of this interaction emphasizes the importance of exploring it further. *Graphics/clipart* searches, however, have half the number of images viewed per preview page visit compared with other query types, but a much higher average dwell time per image. *Celebrity* queries exhibit the shortest overall preview duration and image dwell time, which is consistent with the shorter total session length for these queries. This suggests that users spend more time inspecting images during *graphics/clipart* searches, while they take less time in viewing *celebrity* images possibly in a more casual manner. For *graphics/clipart* and *fashion item* queries, users are also much more likely to click on the referral website. The 'single image previews' column on Table 12 shows the proportion of cases where the user clicks on an image, views the image, and does not preview any other result images. It is largest for *graphics/clipart* queries, which also has the lowest number of images viewed per preview page visit. These results reinforce the query type based differences found in the previous section and, along with the analysis of hover behavior, provide a positive answer to *RQ 2.2*, showing that these additional interaction do, indeed, give further insights into image search behavior.

[9]All 3 major U.S. search engine provide variations of this type of image preview mechanism.

Query Type	Preview Duration (min)	# of Images Viewed	Dwell Time (sec)	Single Image Previews	Referral Page CTR
Celebrities	1.62	18	5.29	35.36%	1
Graphics/Clipart	2.12	9	14.34	49.71%	4.83
Fashion Items	2.32	17	8.38	39.16%	5.67
Animals	2.38	19	7.44	38.23%	1.67
Automobiles	2.17	17	7.50	33.52%	2

Table 12: Preview Page Characteristics by Query Type. (CTR is normalized to hide commercially sensitive data.)

LINKING BEHAVIOR TO SEARCH INTENT

To address *RQ 3 'Can we associate query types with classes of search intent? How does search intent relate to behavior?*, and to support our interpretations of user behavior with the testimonials of real users, we conducted a survey asking users about their most recent image search experience. All respondents, recruited through email, social media, and personal contacts of the authors, fall in the age range of 18 to 41, with 91% between 24 to 35. They completed an online survey (`http://goo.gl/gfO1Uh`), which is a mix of multiple-choice and open-ended questions, including questions specifically related to search intent. The open-ended questions allowed users to freely elaborate on their experience, and were especially useful in giving insights that were not available from the log analysis. To ensure veridicality, we asked the participants to rate, on a 5-pt scale, how confident they were in remembering their experience. We excluded those whose confidence rating was less then 3, as well as those whose last recalled search occurred more than a week previously, leaving 43 participants. As with any qualitative analysis, we acknowledge issues of limited generizability; while a more rigorous survey can further provide a deeper understanding of image search behavior, we believe these initial results are useful for interpreting the log analysis.

To link the user behavior with known classes of image search intent, we use 4 classes of image search intent proposed by Lux et al. [13]: *knowledge orientation*, *navigation*, *mental image*, and *transaction*. Our results suggest a link between the query type based behavior observed in our study and the *knowledge orientation* and *transaction* intent; the other two classes of intent were not referenced by our respondents.

Transaction-Intent Queries. A majority of the respondents stated they were searching for clipart, mostly for presentation slides. They were trying to 'illustrate an idea' using web images or were looking for images 'for further use', which corresponds closely with *transaction* intent [13]. They further responded that they were looking for a *set of images* on *generic* instances. Some signs of transactional intent for *graphics/clipart* searches found during the log-based analysis include: query refinement to improve the results, higher click-through rate compared to other query types, longer dwell time on the preview page (possibly due to image downloads), and higher click-through to the referral website. With regards to the motivation behind clicks and hovers, the main reason for clicking on images while searching for clipart appears to be to view a better-resolution image, or to visit the image webpage. Some argued that hovers were more convenient than clicks: "[Hovering is] similar to clicking but not that exten-

sive". Since *fashion item* queries share many of these behavioral patterns, they may also be associated with transaction intent; in this case though, the user is likely to be viewing images of the items with a view to purchasing them, which would explain the high referral page CTR.

Knowledge Orientation-Intent Searches. Users searching for celebrities answered that they performed the search to 'inform themselves', which corresponds to *knowledge orientation* intent [13]. The log analysis showed that *celebrity* queries lead to relatively shallow interaction with results in the search results and image preview pages, suggesting that browsing the result thumbnails themselves is often enough to satisfy this intent. One respondent, searching for a famous basketball coach, stated that "just looking at the thumbnails (without clicking/hovering) was enough". This suggests that, for *knowledge orientation intent* queries, 'successful' queries may not always result in direct interaction with the results, and explains low click-through rates for this query type. *Celebrity* searchers also responded that they were looking for *a set of images*, which explains why users consume the top ranked images as a set (i.e. rank position was relatively unimportant for CTR). None of those who made celebrity searches responded that they rewrote the query, even though we detected many 'complete change' reformulations in the logs. We posit that complete changes are detected when users search for different celebrities from the initial search within the same session, but the respondents considered this a new query rather than a reformulation. The similar behavior associated with *animal* queries and, to a lesser extent, *automobile* queries, suggests a similar intent.

Is Image Search Exploratory? Previous work on image search log analysis suggested that image search is more exploratory than text search, mainly based on the number of results clicked and the search depth (number of result pages viewed) [1]. One of our research sub-questions, *RQ 2.1*, asked if we could *'find further evidence of exploratory behavior in image search'*. We have shown that *celebrity* queries and *animal* queries, which we are also associated with *knowledge orientation* intent, exhibit many features that suggest exploratory behavior: shallow interaction with the search results and image preview pages, complete query changes to other entities of the same type, more variety in terms of the images clicked/hovered on the results page, and the relative unimportance of rank position for CTR/HTR. The survey feedback also indicated that users were interested in a 'set of images', again indicating their interest in exploring a set of results, rather than finding a single result. *Graphics/clipart* and *fashion item* queries, on the other hand, which we have associated with *transaction* intent, exhibit a number of behavioral features that are more suggestive of a focused search: deeper interaction with the search results page and the image preview page, CTR that is heavily dependent on rank position, lower variety of images clicked and an emphasis on query refinement through partial changes. Andre et al. [1] interpret the *higher* result interaction and *greater* search depth in image search as an indication of image search being more exploratory than web search, while the query types that we interpret as exploratory exhibit *less* result interaction and *shallower* search depth, which would seem to contradict that. In fact, we agree with their interpretation that image search is generally more exploratory than text search; in our survey, a respondent also spoke about looking for *sets* of images for *transaction* intent image queries. We believe that image search has more result interaction than text search because it is inherently more exploratory, due to the large number of relevant images for most queries, and because users want to consume a set of images. Within image search, however, users need to interact with the results of focused queries even more deeply because they have a specific purpose in mind, meaning that it takes more effort to find what they are looking for.

PRACTICAL IMPLICATIONS

Our study has a number of important practical implications. The ways in which the insights from this work can be used to improve search relevance include, but are not limited to, the following:

(1) A query classifier for the query types we identify, along with a matching image classifier, can benefit relevance. For example, if we could classify a query as being a *graphics/clipart* query, and also classify images of the same type, this can be used to improve relevance. Also, the facet-based classification results suggest what types of image classifiers the community should work on (e.g. the *who* facet).

(2) We emphasize 3 alternative sources of implicit relevance feedback from search logs: hovering over images, dwell time in the preview page, and click-through from the image preview to the referral website. We have shown that hovers have a moderate correlation with clicks, which are already known to correlate highly with relevance. Given that there are many more hovers than clicks, this is a promising source of implicit relevance feedback. We also show that for most query types there is an average of at least 17 images viewed on the image preview page for every result click. Given this, interactions with the preview page (e.g. dwell time, referral page CTR) are very promising sources of additional implicit feedback.

(3) We show that CTR/HTR is heavily dependent on query type, as is dwell time and referral page CTR, which has important applications in click modeling [11]; the huge variation in CTR, HTR, dwell time, and referral CTR based on query type shows that, to effectively interpret implicit user feedback as a relevance signal, query type is essential. Similarly, the fact that position information is not always important for CTR/HTR can have important implications in interpreting click/hover feedback.

Overall, we find that only some query types are exploratory, and show which query types can be associated with *transaction* or *knowledge orientation* intent. This could motivate different ranking strategies, or results presentation, based on query type. Specific suggestions are out of scope of this paper, but our results can serve as useful evidence for design decisions or ranking strategies.

SUMMARY & CONCLUSIONS

In this paper we conduct a large-scale study of a web image search log, to understand what users search for and how

their search behavior varies depending on query type. We categorize queries with respect to two orthogonal taxonomies. We show that the head of the query distribution is dominated by *Arts, Culture & Entertainment* queries and by queries for specific entities, mainly people: in other words, *celebrity* queries. The tail of the distribution exhibits much more diversity both in terms of the subject of queries (*fashion items*, *graphics/clipart* and *automobiles*) and their facets/aspects (specific and generic entities or actions). We also show that the following important *query types* emerge at the intersection of subject-based and facet-based taxonomies: *celebrities*, *graphics/clipart*, *fashion items*, *animals*, and *automobiles*.

We show that *celebrity* search sessions and, to a lesser extent, *animal* sessions, are shorter in duration and involve less interaction with search results than other query types, often leading to a browsing-like behavior where the user completely changes the query to a different entity within the same session. We also show that *graphics/clipart* queries in particular, and also *fashion item* queries, show more complex behavior, with longer sessions, more refinement of queries, and deeper interaction with search results. We also show that clickthrough rate (CTR) and image dwell time are much lower for *celebrity* queries than for other query types, and that CTR is much less dependent on the image rank position for *celebrity* and *animal* queries. These query type differences are sometimes very large, which suggests that we should be considering query type when using click logs as user feedback to improve image search relevance.

We further provide possible explanations of these differences in user behavior based on qualitative analysis of individual user experiences in searching for images, reported through a survey. The responses suggest that searches for *graphics/clipart* images are often associated with *transaction* intent, while *celebrity* queries are associated with *knowledge orientation* intent. We also show that the behavioral differences that we observe in our log analysis are consistent with such intent; for *transactional intent* queries, users often have a specific goal in mind that is difficult to articulate, and they take extra effort to view and select images, as well as to refine their queries. Similarly, survey feedback about *celebrity* queries supports our interpretation that such searchs tend to be more casual and exploratory.

For future work, we plan to analyze user behavior variation with respect to facets and aspects from the Shatford-Panofsky taxonomy. Also, while in this study we only deal with search behavior on desktops, it is also important to study the behavior across different device types, and to compare behavior differences between head/tail queries. Finally, it would also be beneficial to evaluate the effectiveness of hovers and preview page interactions for image search relevance.

REFERENCES

1. Andre, P., Cutrell, E., Tan, D. S., and Smith, G. Designing novel image search interfaces by understanding unique characteristics and usage. In *INTERACT 2009*, Springer-Verlag (2009), 340–353.
2. Armitage, L. H., and Enser, P. G. Analysis of user need in image archives. *JIS 23* (1997), 287–299.
3. Broder, A. A taxonomy of web search. *SIGIR Forum 36*, 2 (2002), 3–10.
4. Craswell, N., and Szummer, M. Random walks on the click graph. In *SIGIR 2007*, 239–246.
5. Goodrum, A., and Spink, A. Visual information seeking: A study of image queries on the world wide web. In *ASIS 1999*, vol. 36, ERIC (1999), 665–74.
6. Goodrum, A., and Spink, A. Image searching on the excite web search engine. *IPM 37*, 2 (2001), 295–311.
7. Huang, J., and Efthimiadis, E. N. Analyzing and evaluating query reformulation strategies in web search logs. In *CIKM 2009*, ACM (2009).
8. Jansen, B. J. Searching for digital images on the web. *JDOC 64*, 1 (2008), 81–101.
9. Jansen, B. J., Spink, A., and Narayan, B. Query modifications patterns during web searching. In *ITNG 2007* (2007), 439–444.
10. Jansen, B. J., Spink, A., and Pedersen, J. O. The effect of specialized multimedia collections on web searching. *J. Web Eng. 3*, 3-4 (2004), 182–199.
11. Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR 2005*, 154–161.
12. Liu, J., Cole, M. J., Liu, C., Bierig, R., Gwizdka, J., Belkin, N. J., Zhang, J., and Zhang, X. Search behaviors in different task types. In *JCDL 2010*, 69–78.
13. Lux, M., Kofler, C., and Marques, O. A classification scheme for user intentions in image search. In *Ext. Abstracts CHI 2010*, ACM (2010).
14. Maniu, S., O'Hare, N., Aiello, L., Chiarandini, L., and Jaimes, A. Search behaviour on photo sharing platforms. In *ICME 2013* (2013), 1–6.
15. Ozmutlu, S., Spink, A., and Ozmutlu, H. C. Multimedia web searching trends: 1997–2001. *IPM 39*, 4 (2003), 611–621.
16. Panofsky, E. *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*. Harper & Row, 1972.
17. Pu, H.-T. A comparative analysis of web image and textual queries. *OIR 29*, 5 (2005), 457–467.
18. Shatford, S. Analyzing the subject of a picture: a theoretical approach. *CCQ 6*, 3 (1986), 39–62.
19. Smith, G., Brien, C., and Ashman, H. Evaluating implicit judgments from image search clickthrough data. *JASIST 63*, 12 (Dec. 2012), 2451–2462.
20. Tjondronegoro, D., Spink, A., and Jansen, B. J. A study and comparison of multimedia Web searching: 1997-2006. *JASIST 60* (2009), 1756–1768.
21. Tsikrika, T., and Diou, C. Multi-evidence user group discovery in professional image search. In *ECIR 2014*, Springer (2014), 693–699.