Leveraging User Interaction Signals for Web Image Search

Neil O'Hare
Yahoo Research
Sunnyvale, CA
nohare@yahoo-inc.com

Paloma de Juan
Yahoo
New York, NY
pdjuan@yahoo-inc.com

Rossano Schifanella
University of Turin
Turin, Italy
schifane@di.unito.it

Yunlong He
Yahoo Research
Sunnyvale, CA
yunlong@yahoo-inc.com

Dawei Yin
Yahoo Research
Sunnyvale, CA
daweiy@yahoo-inc.com

Yi Chang
Yahoo Research
Sunnyvale, CA
yichang@yahoo-inc.com

ABSTRACT

User interfaces for web image search engine results differ significantly from interfaces for traditional (text) web search results, supporting a richer interaction. In particular, users can see an enlarged image preview by hovering over a result image, and an 'image preview' page allows users to browse further enlarged versions of the results, and to click-through to the referral page where the image is embedded. No existing work investigates the utility of these interactions as implicit relevance feedback for improving search ranking, beyond using clicks on images displayed in the search results page. In this paper we propose a number of implicit relevance feedback features based on these additional interactions: hover-through rate, 'converted-hover' rate, referral page click through, and a number of dwell time features. Also, since images are never self-contained, but always embedded in a referral page, we posit that clicks on other images that are embedded on the same referral webpage as a given image can carry useful relevance information about that image. We also posit that query-independent versions of implicit feedback features, while not expected to capture topical relevance, will carry feedback about the quality or attractiveness of images, an important dimension of relevance for web image search. In an extensive set of ranking experiments in a learning to rank framework, using a large annotated corpus, the proposed features give statistically significant gains of over 2% compared to a state of the art baseline that uses standard click features.

CCS Concepts

•Information systems → Web search engines; Content ranking; Image search;

Keywords

Web Image Search, User Behavior, Ranking

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2911532

1. INTRODUCTION

Although multimedia search in general, and image search in particular, have been very active research areas for well over a decade, there has been relatively little work in understanding how user behavior, based on interactions with image search engine interfaces, the traces of which are stored in search engine logs, can be used to provide implicit relevance feedback that can improve the rankings for commercial image search engines. Instead, standard approaches, developed with general web search in mind, are applied to multimedia search without any effort to adapt them to the special case of image search and the rich, and unique, user interactions that the interfaces of commercial image search engines facilitate. On the other hand, image search has been shown to be very important within web search: for example, a recent study shows that queries showing image intent were second only to the navigational intent[1] queries on desktops and tablets, and they were the most popular (42%) for mobile phone devices [22].

In this work, we address this gap by leveraging the rich user interactions with web image search results to extract implicit relevance feedback information and improve search engine rankings. Interaction features that are found in image search, but not in general web search, include: (1) 'hovering' over result images to see a preview; (2) browsing enlarged previews of many images after a click on a search result, via next/previous buttons; (3) the ability to further click through, from this image preview, to the host webpage where the image is embedded. In this work, we propose a number of novel implicit user feedback features for image search based on the above interface features, and evaluate their usefulness for image search ranking on a large annotated image search corpus, showing an improvement in relevance over a strong baseline based on click data.

The main contributions of this paper are as follows:

- We propose a 'hover' rate feature based on an image-specific interaction on the *Search Results Page*.
- We propose two sets of user behavior features based on user interactions with the image *Preview Page*: (i) click-through from the preview page to the host/referral page where the image is embedded, and (ii) a set of dwell time features.

[1]navigational queries = queries that seek a single website or web page of a single entity

- We propose propagating the user behavior features beyond the *query-image url* level to the *query-referral page url* level, and also to the *query-independent* level.
- We evaluate these novel features in a learning to rank framework using a large scale annotated corpus, and show that they achieve significant improvements over a strong baseline based on standard click features.

Although other work has explored dwell time, to the best of our knowledge this is the first work to use dwell time for a large scale, general search relevance task.

The rest of the paper is organized as follows: in the next Section we introduce the related work, before going on to describe the unique features of web image search user interfaces in Section 3. In Section 4 we introduce novel implicit relevance feedback features for web image search. In Section 5 we describe the data that we use in our experiments, and then we go on to describe our experiments and results in Section 6. Finally, we conclude the paper in Section 7.

2. RELATED WORK

Click information from the search engine logs has long been used as an implicit relevance feedback signal to improve relevance for web search. Features like click-through rate are calculated from the log data and used as a signal for relevance, under the assumption that relevant documents have more clicks than non-relevant ones [12, 13, 1]. Agichtein et al. [1], for example, show that adding such implicit user behavior information can give as much as a 30% improvement for relevance in web search. Other work tries to better understand user behavior by creating click models that aim to account for biases caused by the display position and other factors [5]. While most of this has focused on web search, and treats the results as a simple ranked list, at least one recent work studied aggregated search on a web search results page, where video and image results, among others, are displayed alongside traditional web results [25].

There has been relatively little work that has aimed at leveraging user behavior information to improve results for multimedia search, either for images or videos. Craswell et al. [4] explore random walk models on the click graph for propagating click information to URLs which have not been clicked. They are not specifically interested in image search, however, but use image data because it has features that suit the research questions on that paper. They show that 75% of clicked images are relevant to the query, which means that basic click data is a very strong indicator of relevance for image search. Jain & Varma [9] use click data to train query-dependent re-ranking models, but do not focus on exploiting click data as a feature for ranking. Smith & Ashman [21] study click-through data for image search and, consistently with Craswell et al., showed that it was 'considerably more accurate in general than document based search click-through data', although the reliability of the clicks was shown to be dependent on factors such as query type and the quality (in terms of precision) of the results shown. Tsikrika *et al.* [24] have used image clicks to automatically create ground truth labels for training visual classifiers, while other work has combined click data and visual features for image ranking [30]. Other researchers have studied web image search interaction logs to understand how interaction with image search engines differs from general search, without exploring the utility of these interactions as implicit relevance information [10, 2, 17, 19].

In traditional web search, beyond standard clicks on the *Search Results Page* (SRP), some researchers have explored post-SRP clicks for improving relevance [3, 26], although this work relies on user-installed browser toolbars to log the post SRP-click user trails, while other work has explored user branching behavior using tabs [8]. Some recent work has investigated using dwell time on search result pages to model user satisfaction [15], although they do not use dwell time as a signal to predict relevance. Other work has explored using dwell time for personalization of web search results [28, 27], while Liu et al. [16] showed that dwell time is a good indicator of document usefulness for a variety of tasks. Yi *et al.* use dwell time as a novel feature in a learning to rank framework for recommendation [29]. With respect to images, Trevisiol et al. [23] leveraged extended user browsing traces on the Flickr photo sharing platform to create a query independent image authority measure.

This work focuses on novel features for implicit relevance feedback beyond the standard SRP clicks. While 'hovers', and clicks from the image *Preview Page* to the image *Referral Page*, are totally unexplored due to their uniqueness to image search, our use of dwell time features is related to the work cited above. We emphasize, however, that the rich interaction with multiple search results *after* an SRP click, but still within the search engine experience (and therefore these interactions are stored in the search engine logs), are unique to image search, providing much more dwell time data than would be available in general web search. Also, we note that we explore propagating click data beyond the *query image url* level, again unique to image search and to this work. To the best of our knowledge, the is the first work to use dwell time features extracted from large scale log data for a search relevance ranking task (although Yi *et al.* [29] did so for a recommendation task).

3. IMAGE SEARCH USER INTERACTION

Image search interfaces differ greatly from traditional web search, supporting a number of interactions not associated with traditional search. Figure 1 (first row) shows the *Search Results Page* for the three major U.S. search engines (Bing, Yahoo and Google). The main section of the layout presents a grid of clickable thumbnails. Two out of the three search engines allow a form of interaction that is specific to image search: the 'hover' (see pointer 1 in Figure 1). After hovering the mouse cursor on a search result image for about 1-2 seconds, the user can see an enlarged version of the thumbnail with some additional information about the context of the image (e.g., the website it was crawled from, the image resolution).

When the user clicks on a thumbnail (or its enlarged version), the *Preview Page* is loaded (see Figure 1, second row). This page is generally displayed as a new view, or it can unfold from the thumbnail. In either case, the *Preview Page* extends over the whole width of the screen, and shows a larger version of the image, not necessarily in the original size. In addition to a link to the actual image (i.e., a full size version hosted in its original server), a *Preview Page* typically provides other elements for interaction:

- **Thumbnails** (see pointer 2 in Figure 1): The user can jump to a specific image by clicking on a thumbnail,

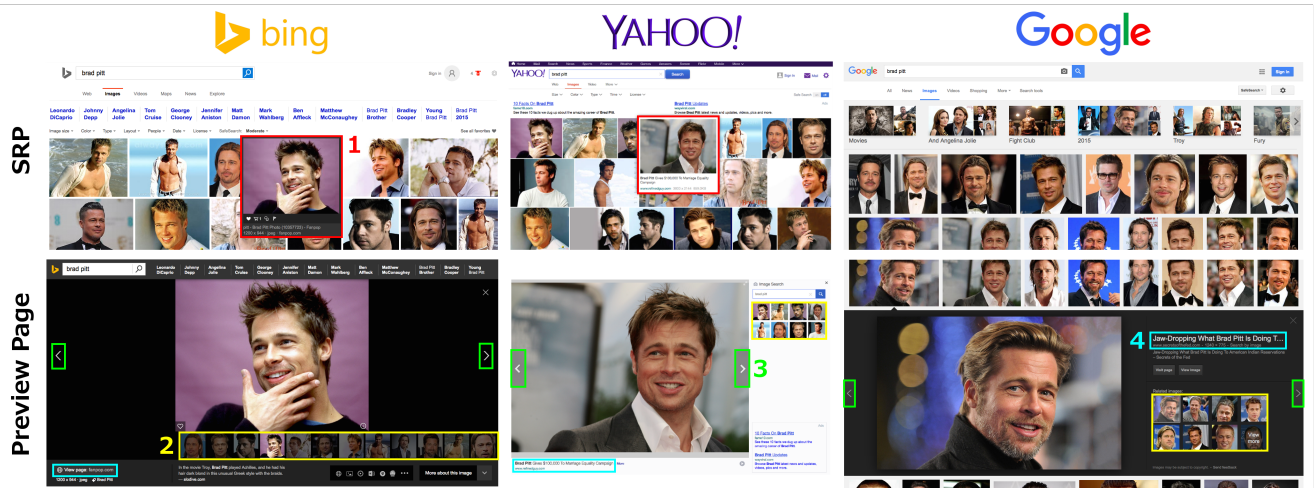

Figure 1: *Search Results Page* (SRP) and *Preview Page* from the three major search engines.

without leaving the *Preview Page*. The thumbnails show neighboring results from the *Search Results Page* (following the original ranking).

- **Navigation** (see pointer 3): By clicking on previous/next buttons, the user can switch to the previous and next images from the thumbnail list without leaving the *Preview Page*.
- **Referral Page** (see pointer 4): Contextual information (title and URL of the *Referral Page* that the image was crawled from, and sometimes a text snippet from the same page) is shown to the user, who can click on a link to go to that *Referral Page*.

The user can additionally 'close' the *Preview Page* to go back to the *Search Results Page*.

3.1 Context of Images in Web

Images are not standalone content items on the web, but instead they are embedded in web pages, that link to the servers where the images are actually hosted. Images are typically discovered by search engine crawlers in the context of the web pages in which they are embedded, and the search engines will store details of that page along with the details of the image. Also, images are not canonical resources pages link to, which means that there is a very high level of replication. Two identical images can be hosted in two different servers (and hence have two different URLs), and any of those locations can be linked to by any number of websites. Additionally, images do not carry any information about what they represent, so search engines need to extract this information from their context, i.e., the surrounding text in the website they are linked from. This means that the information that an image is associated with is heavily dependent on the content of its *Referral Page*.

As the same image (in terms of content or location) can be embedded in different contexts, it is important to take into account the website the image was crawled from when it is shown to the user as a result for a particular query. In other words, much of the information we have available to decide on the relevance of an image to a given query depends heavily on the textual content and metadata of its *Referral Page*, and its power to accurately describe what the image depicts. This means that, when considering implicit user feedback, we should also consider the relevance of the image *referral page* since, when a user interacts with an image result, they are not only validating the relevance of the image to the query, but also indirectly validating the relevance of its *Referral Page* to that same query.

3.2 Relevance Requirements for Image Search

Work on user expectations for relevance in image search in commercial web search engines has shown that, in addition to *topical relevance*, users consider *image quality*, or *attractiveness*, to be of paramount importance. Geng et al [7] have shown that all of the top requirements for image search results are directly related to attractiveness or quality: 'high quality', 'colorful', 'sharp', 'beautiful', 'appealing', 'vivid'. These aspects were actually rated as *more important* than relevance although, as the authors note, this is likely due to the design of the study, where the users were presented with the top results from commercial search engines. In that setting, there were unlikely to be any problems with *topical relevance*, and so these users did not identify this as a major issue. Nevertheless, that work clearly emphasizes that image quality and attractiveness are, alongside topical relevance, the most important user requirements in image search. For this reason, in this paper we will focus our evaluation of image search ranking algorithms on topical relevance and image quality (we use quality in the general sense, to capture objective image quality, beauty and appealingness).

4. IMPLICIT FEEDBACK FEATURES FOR IMAGE SEARCH

In this section we introduce new implicit relevance feedback features for Image Search, first describing features extracted from the *Search Results Page* and the image *Preview Page*, and then describing how we propagate these features to the *Referral Page* level and the *query-independent* level.

4.1 Implicit Feedback Features from the Image Search Results Page

The standard way to exploit user behavior in web search is to use the click-through rate (or click probability) on

the *Search Results Page* (SRP). This is calculated by dividing the number of times an image has been clicked for a given query by the number of times it has been displayed, or viewed, for that query:

$$CTR = \frac{clicks(image\ url, query)}{views(image\ url, query)} \quad (1)$$

In Section 3.1 we also described how some web image search engines also support a 'hover' event, where leaving the mouse hovering over an image for a number of seconds results in an enlarged version of the image being displayed (see Figure 1). Park et al. [19] have shown that image search interfaces elicit 8 to 10 times more hover interactions than clicks, and that the images that are hovered on are moderately to highly correlated with the images that are clicked on. This suggests that, although potentially more noisy than clicks, this noise could be compensated by the larger volume, and hovers could be a strong additional signal for relevance. To exploit the interaction for improving relevance, we propose the hover-through rate (HTR) feature, as follows:

$$HTR = \frac{hovers(image\ url, query)}{views(image\ url, query)} \quad (2)$$

Also, in many sessions, the display of an enlarged 'hover' image may be followed by a standard click on the image, if the user wants to see a bigger version of the image in the *Preview Page*[2]. We can interpret this interaction as the user first inspecting the enlarged images for relevance and, if they judge this image as being relevant, they then click through to the image *Preview Page*. Based on this, we also propose a new feature, which we call *converted hover rate* (CHR), the proportion of hover events that are followed by clicks the same session (i.e. hovers that are 'converted' to clicks):

$$CHR = \frac{converted\ clicks(image\ url, query)}{hovers(image\ url, query)} \quad (3)$$

4.2 Implicit Feedback Features from the Image Preview Page

From the image *Preview Page*, we derive two types of features: (i) features based on clicks from the image *Preview Page* to the original *Referral Page* where the image is hosted, and (ii) features on the dwell time on images in this page. Interactions with this page can be very important, since users often interact with images in this view that they did not click on in the *Search Results Page*: one recent study [19] showed that, for each click on an image from the *Search Results Page* (which opens the *Preview Page*), an average of 17 additional images are viewed by the user in the *Preview Page* via navigation using the previous/next buttons.. Given this, we propose that these interactions carry much additional implicit relevance information.

4.2.1 Referral Page Click-through

All of the major U.S. search engines provide, in the *Image Preview* page, a link to the original *referral page* where the image was crawled from, allowing the user to view the image in its original context. We posit that this deeper method of interaction should be strongly correlated with relevance, and that it can be used as a feature for image relevance. We can calculate the *Referral Page Click-through Rate* (RP-CTR), as follows:

$$RP\text{-}CTR = \frac{referral\ page\ clicks(image\ url, query)}{preview\ page\ views(image\ url, query)} \quad (4)$$

4.2.2 Dwell Time Features

Given that a number of images are consumed on each user visit to the *Preview Page*, we know the time when the user started viewing an image (when they clicked through from the *Search Results Page*, or when they accessed it via the *next* button), and when they finished viewing the image (when they clicked the *next/previous* button, or exiting the *Preview Page*). With this information, we calculate the amount of time that the user spent viewing each image, or the image *dwell time*. Previous work has shown that dwell time can be a strong indicator of user satisfaction for many tasks [16, 15], and it has been used successfully as a feature to improve recommendation [29]. We hypothesize that larger dwell times on a individual images indicates a deeper interaction with results, reflecting higher quality and/or topical relevance. Once we know the dwell time on an individual image, we can calculate a number of features based on it. We calculate three main types of *dwell time* features (see Table 1 for a complete list):

- *Absolute* dwell time features: the dwell time in seconds, and the log of the dwell time.
- *Normalised* dwell time features. Previous work [14] has shown that dwell time can be dependent on both the user and the task. To correct for both of these factors, we normalize the dwell time with respect to other images viewed in the same *Preview Page* session. We define such a session as starting when a user clicks on a result in the *Search Results Page* to enter *Preview Page* mode, and ends when the user exits the image *Preview Page*. This within-session normalization naturally controls for both user and task. Normalised features are calculated by normalizing against the total session duration, the dwell time of previous/subsequent images, etc.
- Binary *categorical* dwell time features calculated relative to the rest of the *preview page* session, such as whether the dwell time is the longest, shortest, greater than the median/mean, less than the median/mean, for that *preview page* session.

To create our final representation of the dwell time features, we calculate the average value of each feature across all sessions.

4.3 Propagating Implicit Feedback Features

Normally, click-through data is calculated at the query-URL level. That is, when a user clicks on a result, the click is associated with the URL of that result. In web image search, this is the URL of the image result. In image search, however, each image is also hosted in the context of the referral webpage where it has been shown, and this webpage can also exhibit relevance to the query; it could be an article or fanpage dedicated to the topic of the query, for example, and may contain multiple images relevant to the query. For this reason, unique to image search, there are two separate

[2]Note that, although these clicks may take place very soon after a hover, the search logging instrumentation is very reliable at successfully logging, and distinguishing, both events.

Feature	Type	Description
Absolute		
DT	Int	Dwell time, in seconds.
$log(DT)$	Float	The log of the dwell time.
Normalised		
$\frac{DT}{\mu}$	Float	Dwell time divided by the mean dwell time (μ) for that session.
$\frac{DT}{median}$	Float	Dwell time divided by the median dwell time for that session.
DT *z-score*	Float	$\frac{DT-\mu}{\sigma}$ where μ is the mean dwell time for the session and σ its standard deviation.
$\frac{DT}{total}$	Float	Dwell time divided by the total dwell time for the session.
$\frac{DT}{min}$	Float	Dwell time divided by the minimum dwell time for any image in the session.
$\frac{DT}{max}$	Float	Dwell time divided by the maximum dwell time for any image in the session.
$\frac{DT}{next}$	Float	Dwell time divided by the dwell time of the next image in the session.
$\frac{DT}{prev}$	Float	Dwell time divided by the dwell time of the previous image in the session.
$\frac{DT}{meanNext}$	Float	Dwell time divided by the mean dwell time of all subsequent images in the session.
$\frac{DT}{medianNext}$	Float	Dwell time divided by the median dwell time of all subsequent images in the session.
$\frac{DT}{meanPrev}$	Float	Dwell time divided by the mean dwell time of all previous images in the session.
$\frac{DT}{medianPrev}$	Float	Dwell time divided by the median dwell time of all previous images in the session.
Categorical		
skipped	Binary	If the dwell time is less than a threshold (0.5 sec), the image is considered to be skipped.
isLast	Binary	The last image viewed in the session.
isSingleImage	Binary	The first and last image viewed in the session.
isMax	Binary	The longest dwell time of all images in the session.
isMin	Binary	The shortest dwell time of all images in the session.
gtMean	Binary	Dwell time is greater than the mean for the session.
ltMean	Binary	Dwell time is less than the mean for the session.
eqMean	Binary	Dwell time is equal to the mean for the session.
gtMedian	Binary	Dwell time is greater than the median for the session.
ltMedian	Binary	Dwell time is less than the median for the session.
eqMedian	Binary	Dwell time is equal to the median for the session.

Table 1: Definition of Dwell time features. Binary features are 1 if true, 0 if false. Features where the denominator is based on previous (or next) images are not defined for the first (or last) image in the session.

URLs that we can associate search result clicks with: the *image URL* and the *referral page URL*. In Sections 4.1 and 4.2 we proposed new features at the *image URL* level. We now propose to propagate these features to the *referral page URL* level or, equivalently, to associate clicks with the *referral page URL* instead of the *image URL*. This should capture referral page relevance information, and also increase coverage of the implicit feedback features for images which have not been clicked, but for which other images from the same *Referral Page* have been clicked. So, in Equations 2, 3 and 4, this simply means replacing $(image\ url, query)$ with *(referral page url, query)*. Similarly, for this version of the dwell time features, we associate the features with the *referral page URL* rather than the *image URL*.

It is also possible to ignore the query completely, and consider the user behavior features at the *query-independent* level. Unlike propagating to the *Referral Page* level, this is not strictly unique to image search. However, we propose that query independent features will capture the *quality* of images, independent of topical relevance. As discussed in Section 3, Geng et al. [7] previously showed that image attractiveness or quality very important for image search. Given this, we posit that propagating click features to the query independent level will capture implicit user feedback about image quality, and so we also calculate all of our proposed features independent of the query.

In summary, we have three levels of propagation for the user behavior features:

- *Q-URL.* The query is associated with the image URL for the clicked/viewed image: this is our default setting and, if not otherwise stated, this version of the implicit feedback features is being used.
- *Q-Refpage.* The query is associated with the URL of the referral page where the image is embedded.
- *Q-Independent.* The features are calculated for the image url, independently of the query.

5. DATA

We take the complete search engine log data, covering all queries across all device types (mobile, desktop, etc) from the Yahoo search engine, covering a 6 month period from

July 2014 to December 2014[3]. From this log, we extract all user interaction events from the *Search Results Page* and the image *Preview Page*.

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 (and their referral pages) displayed on the page. For click events, we have information about the type of click (e.g. click, hover) and the URL of the clicked images and their referral pages. We then aggregate this data and calculate all of the features described in Section 4 for all *query-url* pairs, *query-referral page* pairs, and all *image urls* in the data.

5.1 Corpus for Web Image Search

To create a corpus that we can use to train and evaluate web image search ranking models, we take a traffic based query sample. That is, we calculate the query volume (the number of times the query was issued) for each query in the logs, and sort the queries by volume. Since in this work we are interested in head and torso queries, for which user behavior information is available in the log data, we restrict our corpus to queries from the top 50% of this traffic distribution (i.e. the most popular queries, covering 50% of the overall traffic volume), taking as our sample a total of 9,272 queries from the top 50% of queries.

5.1.1 Relevance Annotations for Image Web Search

The two main facets of relevance for web image search are *topical relevance* and *image quality/attractiveness* [7]. In order to measure each of these facets separately, for candidate image URLs for each query, we gather separate judgements for topical relevance and quality. All of these annotations are carried out by professional editors, trained to use the topical relevance and quality scales outlined below.

For *topical relevance*, we gather judgements on a 3 point-scale:

- *Relevant.* The main subject of the image is the main subject of the query, and is clearly visible.
- *Moderately Relevant.* The image is only partially relevant to the query. Either the subject of the query is fully depicted, but is only the partial focus of the image, or else it is not depicted clearly enough.
- *Non Relevant.* The image fails to match the subject of the query, or matches it so poorly as to be not useful.

For image quality, we gather judgements on the following 5-point scale:

- *Exceptional.* Very appealing images, showing both outstanding professional quality (photographic and/or editing techniques) and high artistic value.
- *Professional.* Professional-quality images (flawless framing, focus, and lightning), which should also be somewhat attractive/appealing.
- *Good.* Standard quality images without technical flaws (subject well framed, in focus, and easily recognizable), and without any artistic value.
- *Fair.* Low quality images with some technical flaws (slightly blurred, slightly over/underexposed, incorrectly framed), which are not very appealing
- *Bad.* Extremely low quality, out of focus, underexposed, badly framed images.

Finally, we create a single combined *relevance + quality* judgement that incorporates both topical relevance and quality, and maps to a standard 5-point PEGFB scale (Perfect, Excellent, Good, Fair, Bad). We do this using a simple heuristic which gives precedence to topical relevance:

- *Non Relevant* images map to *bad*, regardless of quality.
- *Moderately Relevant* images map to *fair*, regardless of quality.
- For *Relevant* images only, the quality score is considered in the final mapping, as follows: $bad \rightarrow bad$, $fair \rightarrow fair$, $good \rightarrow good$, $professional \rightarrow excellent$, $exceptional \rightarrow perfect$.

5.1.2 Image URL Sampling

To avoid biasing our sample towards the features of the images ranked highly by the ranking algorithm used, we take a sample of a 15 images per query from the top 500 results of a baseline ranking algorithm, and have these annotated by editors. After this initial round of annotation, a preliminary analysis showed that some frequent queries had few or no negatively annotated (i.e. moderately/non relevant) images. For these cases, to ensure a good balance of annotations in our corpus, we identified queries for which the initial annotations suffered from this bias, and sampled additional image URLs to be annotated for these queries. This gave a variable number of image URLs annotated per query, resulting in a total of 221,920 annotations over the 9,272 queries in our sample, an average of almost 24 image URLs per query, with the labels (from the combined mapping) distributed as follows: 1,195 *Perfect*, 46,926 *Excellent*, 30,571 *Good*, 67,669 *Fair* and 75,559 *Bad*.

5.2 Feature Coverage in Corpus

Using the corpus described in the previous subsection, Table 2 shows the coverage of each feature in terms of the percentage of judged query-url paris in the corpus for which we have a non-null value, or a positive (i.e., non-zero) value. In order to remove noise from sparse data, when calculating the implicit feedback features, we treat them as undefined if they are based on less than a minimum number of impressions in the *Search Results Page* or the *Preview Page* (we set the threshold at 20). In Table 2, a non-null value indicates that the corresponding feature can be calculated, whereas a null value means there are too few impressions to calculate the feature reliably, so we treat it as undefined (for the feature vectors used in MLR models, undefined features are assigned the value -1). While the *absolute* coverage values are influenced by the amount of log data available for processing and by the method used to sample image urls for each query, we are mostly interested in the *relative* gain in coverage from adding new features and from propagating those features, which should be quite robust against these factors. The coverage of HTR is slightly lower than that of CTR due to the fact that the impressions accounted for the former are desktop-only, since there is no hover action on mobile. The coverage of the Dwell Time (DT) and the RP-CTR features is based on the number of impressions on the *Preview Page*, which must always be preceded by a click on the *Search Results Page*. DT has slightly lower coverage because the dwell time can not be computed for those im-

[3]We do not reveal further details about this initial dataset (e.g. size, etc) as such information is commercially sensitive.

Features	Query-URL		Query-Refpage		Query-Independent	
	Non-null (%)	Non-zero (%)	Non-null (%)	Non-zero (%)	Non-null (%)	Non-zero (%)
CTR	24.93	17.13	29.67	20.37	59.19	36.78
DT	6.35	-	6.96	-	10.40	-
HTR	21.32	**18.28**	26.03	**22.19**	52.58	**43.52**
CHR	7.24	6.64	8.69	7.96	13.93	12.72
RP-CTR	7.05	2.78	8.48	3.24	13.27	6.02
All	24.93	**20.44**	29.68	**24.43**	59.19	**46.51**

Table 2: Feature coverage for implicit feedback features: *non-null* coverage calculated as the proportion of the dataset for which the features can be calculated, *non-zero* calculated as the proportion of the corpus for which the feature has non-zero values.

ages viewed at the end of the session (e.g., if the users closes the browser or navigates away from the *Preview Page*).

The non-zero coverage shows the proportion of query-URLs in the corpus with non-zero values for this metric: that is, there has been at least one click, so this is potentially a positive relevance signal (the non-zero coverage does not apply to the DT features, since by definition the dwell time, if defined, will always be non-zero). Here HTR has a higher coverage than CTR, despite of the fact that it does not include interactions on mobile devices. We can also see that the overall coverage (last row, calculated as proportion of the corpus where at least one of the features is non-zero) increases when we consider all implicit feedback features, and that corpus coverage increases significantly when propagating features. In fact, we see that the coverage increases for all the features with the adoption of the propagation schemes, and it doubles at the query-independent level compared with the query-URL level. This is important because is shows that by adding new features, and propagating them, the number of image URLs that can potentially benefit from these features increases significantly. This also emphasizes how the introduction of image specific user feedback signals can better capture the peculiarities of user behavior in image search, compared with generic web search.

6. EVALUATION

To evaluate our proposed features we adopt a learning to rank framework. We employ the state of the art *GBRank* ranking algorithm [31], which uses a pairwise loss function based on Gradient Boosting Regression Trees (GBDT), as our ranking function. We convert the combined *relevance + quality* judgements to a numeric learning target as follows: Perfect - 4, Excellent -3, Good - 2, Fair - 1, Bad - 0.

We calculate over 2,000 features as input to the learning algorithm to train two separate baseline models, **B1** and **B2**. We evaluate the effectiveness of our proposed features by adding them as additional features on top of each of the baselines, and measure their effect on search ranking. We compare against the first baseline to confirm that our proposed features carry significant relevance information, and against the second to show that our methods can outperform a strong baseline based on standard click features:

- **B1**: The first set of features is purely content-based, as it does not include information about user interactions. The features used for this baseline are:
  - *Query Features*: Query length, query frequency, query category, etc.
  - *Text Matching*: Features based on the *document text* and *anchor text*. For image search, we extract the title and URL of the referral page, keywords and file name from the image URL, the image 'alt' text, the text surrounding the image, and the anchor text. Each of these is indexed into a separate field, and a number of query-URL level text-based similarity scores are calculated for each field. For each of these fields, basic features are computed and then aggregated to form new composite features. The match score can be as simple as a count or can be more complex such as BM25 [20]. Counts include the number of occurrences in the document, the number of missing query terms or the number of extra terms. Other features measure proximity and order of query terms within the document [18].
  - *Referral Page*: We calculate page authority features for the page where the image is embedded using random walk based methods, host level authority, domain level authority, URL authority, etc.
- **B2**: The second baseline includes CTR for query-URL pairs, in addition to all the features from baseline **B1**.

Partition	Queries	Judgements
Training	7,418	177,736
Tuning	463	10,917
Test	1,391	33,267
Total	*9,272*	*221,920*

Table 3: Basic statisistics for the Web Image Search training and evaluation corpus.

We use Normalized Discounted Cumulative Gain (NDCG) as the metric to assess search relevance performance. NDCG has been widely used to assess relevance in the context of search engines [11]. For a ranked list of N documents, Discounted Cumulative Gain (DCG) is calculated as follows:

$$\mathrm{DCG}_N = \sum_{i=1}^{N} \frac{G_i}{\log_2(i+1)}$$

where G_i represents the weight assigned to the label of the document at position i, i.e., 10 for Perfect, 7 for Excellent, 3 for Good, 0.5 for Fair, and 0 for Bad. NDCG normalizes DCG by dividing it by the ideal DCG that would be

Features	NDCG1	NDCG3	NDCG5	NDCG10
B1	0.661	0.664	0.668	0.721
B2	$0.735^{\dagger\dagger}$	$0.731^{\dagger\dagger}$	$0.738^{\dagger\dagger}$	$0.775^{\dagger\dagger}$
B1 + HTR	$0.709^{\dagger\dagger}$	$0.717^{\dagger\dagger}$	$0.727^{\dagger\dagger}$	$0.764^{\dagger\dagger}$
B1 + CHR	0.676	$0.682^{\dagger\dagger}$	$0.685^{\dagger\dagger}$	$0.727^{\dagger}$
B1 + RP-CTR	0.660	0.655	0.664	0.712
B1 + DT	$0.689^{\dagger\dagger}$	$0.687^{\dagger\dagger}$	$0.687^{\dagger\dagger}$	$0.728^{\dagger}$
B2 + HTR	$0.752^{\ddagger}$	$0.738^{\ddagger\ddagger}$	$0.746^{\ddagger\ddagger}$	$0.778^{\ddagger\ddagger}$
B2 + CHR	0.748	$0.738^{\ddagger}$	$0.745^{\ddagger\ddagger}$	$0.778^{\ddagger\ddagger}$
B2 + RP-CTR	0.74	$0.739^{\ddagger}$	$0.747^{\ddagger\ddagger}$	$0.779^{\ddagger\ddagger}$
B2 + DT	$0.745^{\ddagger}$	$0.740^{\ddagger\ddagger}$	$0.746^{\ddagger\ddagger}$	$0.779^{\ddagger\ddagger}$
B2 + All	$0.753^{\ddagger\ddagger}$	$0.737^{\ddagger}$	$0.747^{\ddagger\ddagger}$	$0.779^{\ddagger\ddagger}$
B2 + All (Q-URL + Q-Refpage)	0.747	$0.739^{\ddagger}$	$0.748^{\ddagger\ddagger}$	$0.785^{\ddagger\ddagger}$
B2 + All (Q-URL + Q-Independent)	0.746	$0.739^{\ddagger}$	$0.747^{\ddagger\ddagger}$	$0.781^{\ddagger\ddagger}$
B2 + All (Q-URL + Q-Refpage + Q-Independent)	$\mathbf{0.755^{\ddagger\ddagger}}$	$\mathbf{0.745^{\ddagger\ddagger}}$	$\mathbf{0.754^{\ddagger\ddagger}}$	$\mathbf{0.790^{\ddagger\ddagger}}$

Table 4: Relevance + Quality NDCG Results for Image Search with Implicit Feedback features. †- significantly better than B1. ††- highly significantly better than B1. ‡: significantly better than B2. ‡‡: highly significantly better than B2.

obtained from a perfect ranking: NDCG has the advantage that it is easier to interpret, in that its score always falls between 0 and 1, with 1 indicating a perfect ranking. We use the symbol $NDCG$ to indicate the average of this value over a set of testing queries in our experiments. In the following sections, we will report NDCG1, NDCG3, NDCG5 and NDCG10, referring to NDCG calculated for the top 1 ranked document only (NDCG1), the top 3 ranked documents only (NDCG3), etc. For our main evaluation metric, we calculate NDCG based on the combined *relevance + quality* scale. For a supplemental evaluation, we use the *topical relevance* labels only, on a 3-point scale, where we map *relevant* to Good, *moderately relevant* to Fair, and *non-relevant* to Bad, and calculate NDGC based on this.

Taking the set of annotated queries described in Section 5.1, we partition them into training, tuning and testing sets. The partitioning is query-based (i.e., all judgements for a given query will be assigned to the same partition) as follows: 80% training, 5% tuning, and 15% test. More details of the evaluation corpus can be found in Table 3. The training partition is used to train the models, the tuning partition is used to conduct a parameter search for the optimal parameters of the *GBRank* algorithm (i.e., the parameter set with the best NDCG score on the tuning partition is selected), and then the model trained with these optimal parameters is tested against the test set to produce our evaluation results.

6.1 Results

6.1.1 Relevance + Quality NDCG

The NDCG results for *Relevance+Quality* are shown in Table 4. Firstly, we can see that all of the proposed features, except for the RP-CTR, give significant improvements over **B1**. For NDCG5, for example, this improvement is 8.2% for the HTR, 2.5% for the CHR, and 2.8% for the DT, and all of these differences are highly statistically significant ($p < 0.01$). The B2 + HTR model outperforms B2 by over 2% for NDCG1, showing a highly significant improvement over B2 for NDCG3, NDCG5 and NDCG10. In fact, all of the proposed new features show at least significant improvements ($p < 0.05$) over B2 of around 1% for NDCG3, highly significant improvements of around 1% for NDCG5, and smaller, but highly significant, improvements for NDCG10. It is also noteworthy that the RP-CTR, which did not improve over **B1**, achieves improvements over **B2** that are comparable with the other proposed features. We hypothesize that the lack of coverage for this feature means that it is not a reliable signal for relevance in the absence of other click data, but since it provides complementary information to *Search Results Page* clicks, this feature can improve relevance in combination with other features. We can also see that the combination of all features gives the best overall performance, even though the improvement gained by adding each of the individual features to **B2** is relatively minor.

These results show that the proposed features can give significant improvements over the strong, state of the art baseline (B2) although the improvements are relatively small at approximately 1% or less for most metrics, with the exception of NDCG1 which shows an improvement of over 2%. Looking at the results for the propagated features, we get further improvements, with all metrics showing highly significant improvements over **B2** of 2% or greater. This confirms that propagating the click features also gives important improvements for image search relevance, and the fact that these improvements can be seen at a greater ranking depth

Features	NDCG1	NDCG5	NDCG10
B2	0.735	0.738	0.775
With Hover Features			
B2 + All	$0.753^{\ddagger\ddagger}$	$0.747^{\ddagger\ddagger}$	$0.779^{\ddagger}$
B2 + All (propagated)	$\mathbf{0.755^{\ddagger\ddagger}}$	$\mathbf{0.754^{\ddagger\ddagger}}$	$\mathbf{0.790^{\ddagger\ddagger}}$
Without Hover Features			
B2 + All	0.747	$0.748^{\ddagger\ddagger}$	$0.780^{\ddagger\ddagger}$
B2 + All (propagated)	$0.752^{\ddagger}$	$0.752^{\ddagger\ddagger}$	$0.788^{\ddagger\ddagger}$

Table 5: Relevance + Quality NDCG Results for Image Search with and without hover-based Implicit Feedback features. †- significantly better than B1. ††- highly significantly better than B1. ‡: significantly better than B2. ‡‡: highly significantly better than B2.

(NDCG10) also shows that these features will give more robust ranking at depth, which is important in image search, where users explore search results much more deeply than in web search, often examining several pages of results [2, 19].

Relative importance of dwell time features.

To understand which among the 25 proposed dwell time features are most useful, we take the relative importance score of each feature in the best *B1 + DT* model, which is calculated by accumulating decrease of loss of each tree splitting point [6]. From this, we find that the following, in order of importance, are the most important dwell time features for this model: *gtMedian*, *DT z-score*, $DT/Mean$, $DT/Total$, $DT/Median$. Four of these features are normalised features, and one is a categorical features. These results seem to confirm that our proposed normalized and categorical dwell time features are important if we are to make the best use of dwell time information.

6.1.2 Topical Relevance only NDCG

We use NDCG for *Relevance+Quality* as our main evaluation metric because we believe that it captures topical relevance while also boosting the scores for methods that give higher rank to the most attractive, high-quality images. However, to verify that the improvements in quality do not come at the expense of topical relevance, we also report, in Table 6, results for topical relevance only NDCG. The results in Table 6 show very similar improvements over the baselines: the best method shows around 1.5% improvement for NDCG1, NDCG5 and NDCG10, and this improvement is highly significant for NDCG5 and NDCG10. These results demonstrate that, whether we focus on *topical relevance* only, or *relevance + quality*, our proposed features for image search ranking give non-trivial, and highly significant improvements over a strong baseline.

Features	NDCG1	NDCG5	NDCG10
B1	0.818	0.782	0.785
B2	$0.893^{\dagger\dagger}$	$0.850^{\dagger\dagger}$	$0.836^{\dagger\dagger}$
B2 + All	0.905	$0.857^{\ddagger\ddagger}$	$0.839^{\ddagger\ddagger}$
B2 + All (propagated)	0.906	$\mathbf{0.862}^{\ddagger\ddagger}$	$\mathbf{0.849}^{\ddagger\ddagger}$

Table 6: Topical Relevance NDCG Results for Image Search with Implicit Feedback features. †- significantly better than B1. ††- highly significantly better than B1. ‡: significantly better than B2. ‡‡: highly significantly better than B2.

6.1.3 Removing the Hover Features

At the time of the study, two of the three major U.S search engines had a 'hover' interaction in their web image search SRP interface. Since then, one of these removed this feature, meaning that only one of these three search engine now supports this interaction, which raises questions as to whether the results reported so far in this paper are generalizable, or whether they are dependent on transient user interface features. The other web image search user interaction features highlighted in Section 3 on which our proposed features are based, however, have been stable features of all of these search engines for a number of years now, and this can be expected to continue. Table 5 shows results for our methods without the hover-based features, as the results without these features are based on more stable, and therefore more generalizable, features. We can see that removing the hover-based features has a relatively small impact on the NDCG; the improvement over the baseline is still more than 2% (and statistically significant) better than **B2** for NDCG1, and it is highly significantly, and over 1.5% better, for NDCG5 and NDCG10.

7. CONCLUSIONS

In this paper we have proposed a number of novel user behavior features for improving web image search ranking: (1) hover-through rate based on 'hovers' in the *Search Results Page*; (2) click-through from the image *Preview Page* to the *Referral Page* where the images are embedded; (3) dwell time features from the image *Preview Page*; and (4) propagating user behavior features to the query-referral page and the query-independent level. We created a large scale corpus to evaluate the usefulness of these features in a learning to rank framework, and show that the new features achieve highly statistically significant improvements of over 2% beyond a strong, state of the art baseline based on standard click-based features.

While this study focused on image search, many of the user interactions discussed here are shared by web video search interfaces, and so we plan to explore the usefulness of the same features for video search. Also, the latest image search interfaces now support clicking through from the *Preview Page* directly to a full size view of the image. This was not available at the time of our study, but we expect that it will provide similar relevance information as click through from the *Preview Page* to the *Referral Page*, and plan to explore this in the future.

8. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 19–26, New York, NY, USA, 2006. ACM.

[2] P. Andre, E. Cutrell, D. S. Tan, and G. Smith. Designing novel image search interfaces by understanding unique characteristics and usage. In *INTERACT 2009*, pages 340–353. Springer-Verlag, 2009.

[3] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: Identifying relevant websites from user activity. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 51–60, New York, NY, USA, 2008. ACM.

[4] N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 239–246, New York, NY, USA, 2007. ACM.

[5] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 87–94, New York, NY, USA, 2008. ACM.

[6] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[7] B. Geng, L. Yang, C. Xu, X.-S. Hua, and S. Li. The role of attractiveness in web image search. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, pages 63–72, New York, NY, USA, 2011. ACM.

[8] J. Huang, T. Lin, and R. W. White. No search result left behind: branching behavior with browser tabs. In *WSDM*, 2012.

[9] V. Jain and M. Varma. Learning to re-rank: Query-dependent image re-ranking using click data. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 277–286, New York, NY, USA, 2011. ACM.

[10] B. J. Jansen, A. Spink, and J. O. Pedersen. The effect of specialized multimedia collections on web searching. *J. Web Eng.*, 3(3-4):182–199, 2004.

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM TOIS*.

[12] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 154–161, New York, NY, USA, 2005. ACM.

[14] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 377–384, New York, NY, USA, 2004. ACM.

[15] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 193–202, New York, NY, USA, 2014. ACM.

[16] C. Liu, J. Liu, N. Belkin, M. Cole, and J. Gwizdka. Using dwell time as an implicit measure of usefulness in different task types. *Proceedings of the American Society for Information Science and Technology*, 2011.

[17] S. Maniu, N. O'Hare, L. Aiello, L. Chiarandini, and A. Jaimes. Search behaviour on photo sharing platforms. In *ICME 2013*, pages 1–6, 2013.

[18] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 472–479, New York, NY, USA, 2005. ACM.

[19] J. Y. Park, N. O'Hare, R. Schifanella, A. Jaimes, and C.-W. Chung. A large-scale study of user image search behavior on the web. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 985–994, New York, NY, USA, 2015. ACM.

[20] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, Apr. 2009.

[21] G. Smith and H. Ashman. Evaluating implicit judgements from Image search interactions. In *Proceedings of the WebSci'09: Society On-Line*, Mar. 2009.

[22] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1201–1212, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[23] M. Trevisiol, L. Chiarandini, L. M. Aiello, and A. Jaimes. Image ranking based on user browsing behavior. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 445–454, New York, NY, USA, 2012. ACM.

[24] T. Tsikrika, C. Diou, A. P. de Vries, and A. Delopoulos. Image annotation using clickthrough data. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 14:1–14:8, New York, NY, USA, 2009. ACM.

[25] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 503–512, New York, NY, USA, 2013. ACM.

[26] R. W. White and J. Huang. Assessing the scenic route: Measuring the value of search trails in web logs. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 587–594, New York, NY, USA, 2010. ACM.

[27] S. Xu, H. Jiang, and F. C. M. Lau. Mining user dwell time for personalized web search re-ranking. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2367–2372. AAAI Press, 2011.

[28] S. Xu, Y. Zhu, H. Jiang, and F. C. M. Lau. A user-oriented webpage ranking algorithm based on user attention time. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 1255–1260, 2008.

[29] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan. Beyond clicks: Dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 113–120, New York, NY, USA, 2014. ACM.

[30] J. Yu, D. Tao, M. Wang, and Y. Rui. Learning to rank using user clicks and visual features for image retrieval. *IEEE Trans. Cybernetics*, 45(4):767–779, 2015.

[31] Z. Zheng, K. Chen, G. Sun, and H. Zha. A regression framework for learning ranking functions using relative relevance judgments. In *Proceedings of SIGIR '07*.