# From Queries to Cards

## Re-ranking Proactive Card Recommendations Based on Reactive Search History

Milad Shokouhi
Microsoft
Cambridge, United Kingdom
milads@microsoft.com

Qi Guo
Microsoft
Bellevue, Washington
qiguo@microsoft.com

## General Terms

The growing accessibility of mobile devices has substantially reformed the way users access information. While the *reactive* search by query remains as common as before, recent years have witnessed the emergence of various *proactive* systems such as Google Now and Microsoft Cortana. In these systems, relevant content is presented to users based on their context without a query. Interestingly, despite the increasing popularity of such services, there is very little known about how users interact with them.

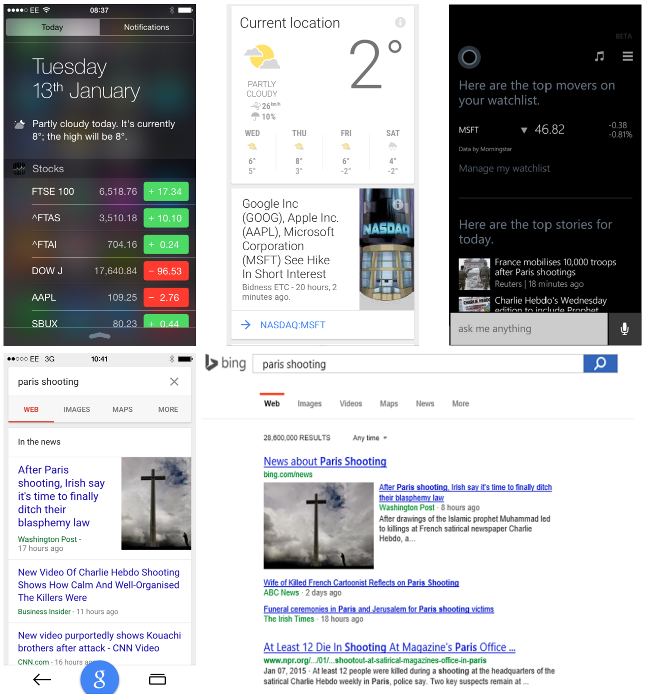
In this paper, we present the first study on user interactions with information cards. We demonstrate that the usage patterns of these cards vary depending on time and location. We also show that while overall different topics are clicked by users on proactive and reactive *platforms*, the topics of the clicked documents by the same user tend to be consistent cross-platform. Furthermore, we propose a supervised framework for re-ranking proactive cards based on the user's context and past history. To train our models, we use the viewport duration and clicks to infer pseudo-relevance labels for the cards. Our results suggest that the quality of card ranking can be significantly improved particularly when the user's reactive search history is matched against the proactive data about the cards.

## Categories and Subject Descriptors

Information Systems [**Information Retrieval**]: Users and interactive retrieval—*Personalization*

## Keywords

Proactive ranking; zero query; Information cards



●●●●○ EE 08:37
Today | Notifications
Tuesday
13th January
Partly cloudy today. It's currently 8°; the high will be 8°.
Stocks

| | | |
|---|---|---|
| FTSE 100 | 6,518.76 | + 17.34 |
| ^FTAS | 3,510.18 | + 10.10 |
| ^FTAI | 704.16 | + 0.24 |
| DOW J | 17,640.84 | – 96.53 |
| AAPL | 109.25 | – 2.76 |
| SBUX | 80.23 | + 0.44 |

Current location
PARTLY CLOUDY
26 km/h
10%
2°

| WED | THU | FRI | SAT |
|---|---|---|---|
| 6°<br>5° | 8°<br>3° | 6°<br>-2° | 4°<br>-2° |

Google Inc (GOOG), Apple Inc. (AAPL), Microsoft Corporation (MSFT) See Hike In Short Interest
Bidness ETC - 20 hours, 2 minutes ago.
NASDAQ:MSFT

BETA
Here are the top movers on your watchlist.
MSFT ▼ 46.82 -0.38 -0.81%
Data by Morningstar
Manage my watchlist
Here are the top stories for today.
France mobilises 10,000 troops after Paris shootings
Reuters | 18 minutes ago
Charlie Hebdo's Wednesday edition to include Prophet
ask me anything

●●○○○ EE 3G 10:41
paris shooting
WEB IMAGES MAPS MORE
In the news
After Paris shooting, Irish say it's time to finally ditch their blasphemy law
Washington Post - 17 hours ago
New Video Of Charlie Hebdo Shooting Shows How Calm And Well-Organised The Killers Were
Business Insider - 11 hours ago
New video purportedly shows Kouachi brothers after attack - CNN Video
CNN.com - 16 hours ago

bing paris shooting
Web Images Videos Maps News More
28,600,000 RESULTS Any time
News about Paris Shooting
bing.com/news
After Paris shooting, Irish say it's time to finally ditch their blasphemy law
Washington Post · 8 hours ago
After drawings of the Islamic prophet Muhammad led to killings at French satirical newspaper Charlie Hebdo, a...
Wife of Killed French Cartoonist Reflects on Paris Shooting
ABC News · 2 days ago
Funeral ceremonies in Paris and Jerusalem for Paris shooting victims
The Irish Times · 18 hours ago
At Least 12 Die In Shooting At Magazine's Paris Office ...
www.npr.org/.../01/...shootout-at-satirical-magazines-office-in-paris
Jan 07, 2015 · At least 12 people were killed during a shooting at the headquarters of the satirical Charlie Hebdo weekly in Paris, police say. Two key suspects remain at ...

Figure 1: Examples of interfaces with *proactive* "cards" presented respectively in Apple Siri, Google Now and Microsoft Cortana (top). Examples of *reactive* search queries submitted to Google on mobile and Bing on desktop (bottom) for query `paris shooting`.

# 1. INTRODUCTION

Mobile devices account for a significant fraction of online search and browsing traffic. The number of mobile queries has grown fivefold in the past three years [6]. In fact, the number of mobile queries has recently exceeded the number of those submitted from desktop devices in the United States and many other countries [1]. Consequently, information retrieval systems are constantly evolving into more contextual and mobile-friendly services.

Among recent enhancements, the introduction of *zero-query* systems [8] is arguably one of the most significant breakthroughs. In 2012 at a workshop, a group of 45 leading experts in the field acknowledged zero-query search to be one of the six most interesting research directions in information retrieval (IR).

> "Future information retrieval systems must anticipate user needs and respond with information appropriate to the current context without the user having to enter a query." [8]

The major players in the search industry have embraced the new trends by releasing systems such as Google Now [4], Apple Siri [5] and Microsoft Cortana [2]. In all these systems, in addition to the typical *reactive search* by a query, *proactive information cards* are presented to users based on their context. Proactive systems are

*SIGIR'15,* August 09 - 13, 2015, Santiago, Chile.
© 2015 ACM. ISBN 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767705.

typically implemented in form of *apps* that once triggered by the user, present several information cards. Each card is particularly designed to satisfy a domain-specific set of information needs. Many of these apps can also run in the background and pop up certain cards proactively depending on the user's context. The top screenshots in Figure 1 include a few examples of proactive cards from Apple Siri (weather, stocks), Google Now (weather, news) and Microsoft Cortana (stock, news) respectively from left to right. It is worth pointing out that there are no explicit queries associated with these cards. In comparison, the screenshots at the bottom of Figure 1 depict two typical reactive searches on – Google – mobile (left), and – Bing – desktop (right) for query `paris shooting`.

Despite the growing importance and popularity of proactive systems, not much have been published on how users interact with them. In this paper, we present the first study of such kind. We start by investigating various aspects of user engagement with different types of cards. For instance, we show that users are likely to interact with different types of cards depending on the time of the day, and to a lesser degree based on their location. In addition, we demonstrate that while the overall topical distributions of clicked documents differ in reactive and proactive logs, the topics of documents that are clicked by the same user tend to be similar between proactive and reactive interactions.

In the light of observations made in the first part of the paper, we focus on improving the quality of proactive ranking in the second part. We propose Carré, a card reranking model that deploys several features based on the user's context and history to improve the ranking of proactive recommendations. Our experimental results conducted over large-scale logs of a commercial search engine confirm that the reactive search history of users – along with several other contextual features – can be used to improve the ranking of proactive recommendations.

The key research contributions of this paper are fourfold,

- Present the first study of user engagement with proactive cards. In particular, we explore how presentation position, local and temporal aspects can affect the user engagement with cards (Section 4).
- Characterize the cross-platform clicks of users and explore the topical similarity of clicked documents across platforms (Section 5).
- Propose Carré, a card reranking model based on context and past history. Inspired by previous work on search personalization [13, 22], we devise an offline optimization framework for reranking proactive cards based on context (Section 6).
- Evaluate card reranking models by inferring pseudo-relevance labels not only based on clicks, but also by viewport tracking of their mobile screens (Section 7).

The remainder of this paper is structured as follows. We continue by summarizing the most related work in the next section. In Section 3, we define our terminology and explain the user logs that are used in our experiments. We study the card usage patterns with respect to presentation position, time and location in Section 4. We characterize the cross-platform clicks in Section 5 and show that the topics of interest for a user are typically consistent across platforms. We then introduce a supervised model for reranking cards based on the user's context and past history in Section 6, and evaluate its performance in Section 7. We summarize our findings and discuss their limitations in Section 8, and conclude in Section 9.

## 2. RELATED WORK

While the study of proactive systems, as introduced in this paper, is a new research area, it has roots in earlier research. A proactive system *aggregates* information from multiple sources (e.g., weather, news, sports) as *recommendations* in the form of *cards*, according to *personal* preferences. Hence, our work is related to the areas of aggregated search, information cards, personalization, and recommendations. In this paper, we also focus on predicting future interactions with the cards for this new modality, which is also related to prior work on click prediction. Next, we will discuss the related research in each of the aforementioned areas.

*Proactive information retrieval.* Rhodes and Maes [37] introduced just in-time information retrieval (JITIR) agents as softwares that proactively retrieve and present relevant information based on the user's context. The context could be inferred based on the content of documents viewed by the user, or physical attributes such as time and location. As in earlier systems [14, 31], Rhodes and Maes [37] focused on textual recommendations such as keywords, emails and documents. However, the content presented in modern information cards is substantially more heterogeneous. The authors nonetheless listed three key features for JITIR namely, (1) proactive (2) non-intrusive, and (3) contextual. that are all still critical in the context of modern proactive systems.

*Aggregated search.* The proactive system is an aggregator of various types of information that are valuable to the user, such as calendar, weather, news, stocks, and places. Its counterpart for the reactive scenario is the aggregated search, where contents from specialized verticals are blended into organic web search results. Diaz [21] studied the integration of news vertical into web search, and proposed modeling the collection and query dynamics to detect queries with the news intent. Arguello et al. [9] later extended this line of research to integrate various types of verticals, such as news, travel, images, and videos. Ponnuswami et al. [34] proposed a randomized online audition framework for optimizing the aggregated ranking based on clicks.

*Information cards.* The cards presented proactively do not always need to be clicked to be useful. Sometimes, the displayed content of the card is sufficient to satisfy the user's information need. This is analogous to *good abandonment* [30] in aggregated search, where the aggregated content from verticals do not always need to be clicked to satisfy the user's intent. Recent research by Guo et al. [24] shed some light on this problem on mobile devices. The authors modeled viewing behavior based on touch interactions, and demonstrated the correlation of document relevance and viewport changing patterns on touch-enabled mobile devices. Lagun et al. [29] extended this line of research to model the viewports for inferring user attention and satisfaction on the reactive search engine result pages. In this paper, we incorporate the viewport-based dwell time as card (pseudo) relevance labels to address this long-standing challenge of "good abandonment".

*Personalization.* Proactive recommendations are highly personalized based on the user interests and context, thus making personalization an integral part the system. In reactive web search, personalization has been well studied for tailoring search results by modeling individual user preferences from history interactions. One of earlier works on leveraging user interaction to infer document relevance is by Fox et al. [22], who showed that the user behavior such as dwell time, has strong association with explicit judgments of satisfaction and can be served as implicit feedback to improve the search experience. Following this research, Xu et al. [48], proposed a method to estimate personalized word-level dwell time from the user's search and browse history. The proposed model can be used to estimate personalized dwell time on new documents to customize search result ranking. More recently, White et al. [46] proposed to model cohorts of users who conduct similar tasks to improve the search personalization.

*Recommender systems.* The proactive system also shares many properties of typical recommender systems [36]. Similar to such systems, we aim to present the most relevant content to the user in the absence of an explicit query. However, the recommendations in the proactive system are for a smaller set of items that are highly repeatable and heterogeneous, and need to be personalized and contextualized in its ranking. Furthermore, we do not compare preferences across users and items as is common in collaborative filtering systems [40].

*Click prediction.* Different models have been proposed in previous research to address various biases in presentation which affect the prediction of future click-through. One example is by Chapelle and Zhang [19], in which the authors proposed a Dynamic Bayesian Network model, aiming to recover actual relevance from clicks by separating the modeling of presentation bias in result attractiveness. Our task are similar in that we aim to predict future interactions with cards presented by the proactive system. However, the major differences lie in the lack of explicit queries, leading to heavier contextualized modeling of the user interests, and alleviating the lack of clicks by utilizing viewport-based metrics.

We will revisit this list in the following sections and provide more details about the most relevant studies when appropriate.

## 3. TERMINOLOGY & DATA

In this section, we introduce the data and the terminology that we use in the rest of of this paper.

*Terminology.* We follow the common IR terminology when referring to reactive searches. Each reactive search *impression* consists of a query submitted by a user at a given time and location, together with all clicks and interaction signals that were collected on those results. Clicks followed by a dwell time of 30 seconds or longer are considered as satisfied (SAT) clicks.[1] As in previous work [17, 35, 45], a *session* is defined as a sequence of impressions with no interval of inactivity longer than 30 minutes.

For proactive scenarios we borrow the same terminology and apply similar constraints. A proactive impression consists of a ranking of cards presented to the user together with corresponding interaction logs recorded such as clicks, viewports and scrolling. Viewport tracking is enabled through JavaScript embedded in the proactive impressions, and the viewport data is buffered and then sent back to the server through HTTP requests. We record the screen size, the positions and sizes of the cards rendered on the proactive impressions in pixels, as well as the viewport changing events with timestamps, allowing us to reconstruct the viewing behavior of the users and calculate the time users spent dwelling on each card.

While reactive impressions start with a query, a proactive impression is triggered when the user launches the app (e.g. clicking on Google Now, or Cortana icons). Consistent with the proactive scenario, a session is comprised of all user interactions with the cards with no interval of inactivity longer than half an hour. In our *cross-platform* analysis, we monitor the behavior of the same users in reactive and proactive sessions.

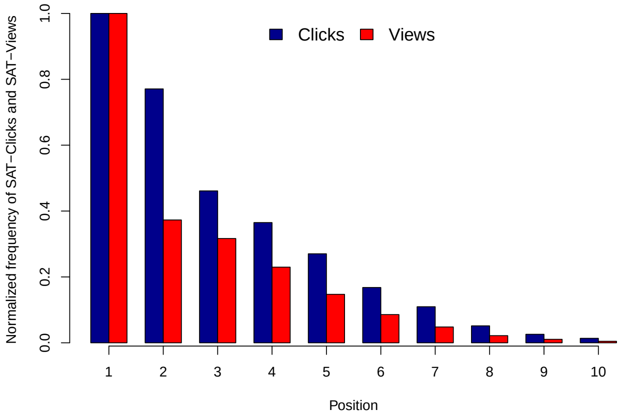


Figure 2: The relative SAT-click and SAT-view frequency recorded across different positions. The numbers are re-scaled with respect to frequency at the first position.

*User logs.* For our experiments we first sampled 365,612 unique users randomly from the proactive logs of Microsoft Cortana [2]. We then collected all proactive history of these users between 08 Nov 2014 – 7 Dec 2014 (1 month). In total, there are 2,663,472 proactive impressions in our dataset. We use the first week of data to study the general card usage patterns (Section 4). The remaining impressions from the last three weeks (811,681 impressions in total after removing impressions with no engagement[2]) are used to train and evaluate our card reranking models (Section 6).

We also extracted the reactive history of this set of users from the query logs of Bing search engine between 1 Sep 2014 – 31 Oct 2014. We deliberately chose an earlier non-overlapping period to avoid any potential patterns that might be caused by general temporal trends rather than user related factors. Each user is distinguished by a unique and anonymized identifier (based on Microsoft Live ID) which is common across Bing and Cortana platforms. This *user-ID* is persistent across both sets of logs and can be used to join all related information about a user. Therefore, the reactive search history of users can be easily compared against their proactive activities. In total, there are 4,113,977 reactive (search) impressions in our dataset.

## 4. PROACTIVE INTERACTIONS

In order to learn effective models for ranking cards, it is essential to understand how users interact with them. While reactive query logs have been extensively studied [10, 27, 28], to the best of our knowledge, there is no previous published work on proactive cards. This section aims to analyze the user interaction with proactive cards and verify if similar patterns hold in general. We later use the insights obtained from these analyses to generate effective features for ranking cards.

For the experiments in this section, we only report the results for seven types of cards namely: sports, places, calendar, finance, news, traffic, and weather. These are the most popular cards in terms of user interactions, and represent more than 80% of proactive *engagement* (clicks, views) in our dataset. The *sports* card, shows

[1] In this paper, clicks always mean SAT clicks regardless of whether they are on card or on a document.

[2] SAT click or long dwell on viewport.

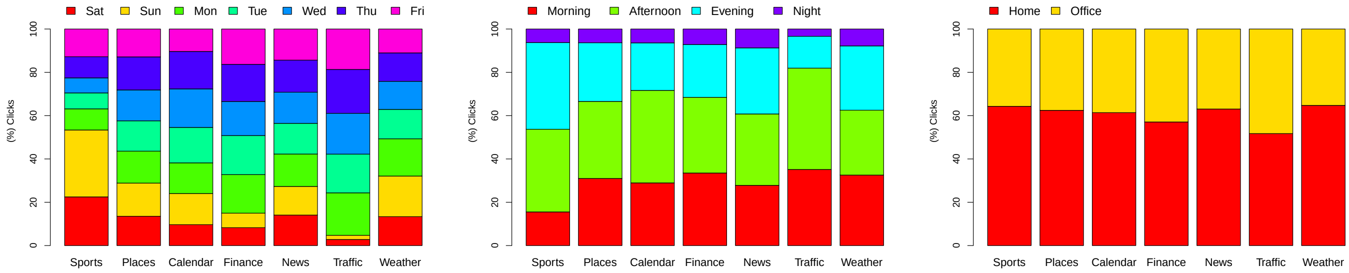


Figure 3: The percentage of clicks on each card given the day of the week (left) and time of day (middle). The percentage of clicks on each card given the location (home vs. office) of the user (right).

the latest scores and fixtures for the team(s) specified by the user. The *places* card recommends restaurants and points of interests nearby based on the user's current location as identified by GPS. The *calendar* card lists the upcoming events and appointments, while the *finance* card tracks the latest stock values for the corporation(s) pre-specified by the user. The *news* card presents the most recent headlines, while the *weather* and *traffic* cards provide forecasts and updates based on the user's current location and most frequently visited destinations (e.g. home and work).

*The effect of card position.* The click distributions on different positions of search results have been well-studied in the past. Agichtein et al. [7] showed that the click distribution on search results is significantly *top-heavy* where the top-ranked documents account for most of the clicks. Joachims et al. [28] explained the skewed distribution by associating it to both relevance and *position bias*. Similar trends have been reported on other ranking platforms such as auto-completion [26] and mobile search results [29]. In the latter case, it is worth pointing out that the authors mostly focused on gaze and viewport rather than clicks. Interestingly, they reported trends that, contrary to previous studies, were not monotonically decreasing by position.

In Figure 2 we repeat a similar analysis on cards. For now, please focus on the blue bars and ignore the red ones as we will come back to them later. Here, the x-axis corresponds to the card position (in the ranking of cards presented to the user), and the y-axis shows the relative click frequency on that position. The frequencies are normalized so that the click frequency at the top position is equal to 1. The trends are consistent with those reported in the past for search queries [7, 28]. Most of the clicks are recorded on the top positions and cards presented at lower ranks get relatively little engagement. However, further investigation would be needed to distinguish the impact of position bias from relevance.

*Temporal & local trends.* The user interests and engagement are affected by various factors. For instance, Beitzel et al. [10] showed that users are likely to query for different topics depending on the time of the day. Similarly, Halvey et al. [25] reported that users web surfing patterns and browsed documents vary depending on time. We verify if the same patterns exist for proactive cards. Figure 3 (left) depicts the percentage of clicks on each card on different days of the week. The boxes in each column are color coded according to the day of the week and their total value sums to one. It is immediately apparent that the card usage varies depending on the day. For example, the traffic card is used on weekdays most often, while the sports card is more likely to be used during weekends. The news and weather cards have roughly uniform usage throughout the week, while the finance card has relatively little engagement during the weekends. This is consistent with findings of Mei and Church [32] on reactive logs that suggested users issue more entertainment-related and fewer business-related queries during weekends.

The middle plot in Figure 3 provides a similar breakdown but for time of the day. We divide the day into four parts based on the time as follows: Morning (6-12], Afternoon (12-18], Evening (18-24], and Night (24-6]. It turns out that users are more likely to click on traffic and finance cards in the morning. The latter is consistent with the observations of Beitzel et al. [10] on search logs, that suggested users are more likely to issue finance-related queries in the morning. Among the cards, news and sports have highest engagement during evening hours whereas the calendar and traffic cards are used more often in the afternoon. Note that the ranking of cards presented to each user does not vary significantly over time in our datasets. On average, more than 82% of cards presented in two consecutive impressions are the same (usually with updated content). The position of a card in the ranking is also fairly stable. On average, the rank of a card changes by *half* a position in two consecutive proactive impressions. The overall trends do not substantially vary either. For instance, the finance and sport cards respectively appear in 2.5%–2.6% and 2.5%–4.9% of daily impressions, although as we observed, their click likelihood could be very different on each day. Therefore, we are confident that the trends observed here are mostly due to temporal factors.

Similar to time, the impact of user location on search behavior has been the subject of extensive research [12, 38, 44]. Cho et al. [20] reported that the mobility of users is mostly centered around two locations: home and work. Therefore, we devote our focus to card usage patterns at work and at home. We infer the centroid of the GPS data from the user in the evening as home, and the centroid of the GPS data during the day as office. Further details of our location inference models are beyond the scope of this paper, but similar approaches can be found elsewhere [20]. We consider the area within 1*km* radius of the user's home location as home area. The office area is defined similarly based on the inferred work location of the user. The right plot in Figure 3 illustrates the percentage of card usage split between home and work. The results suggest that – ignoring clicks from miscellaneous locations – two-third of card clicks take place at home and one-third at the office. Overall, there seems to be very little difference between cards in terms of their usage patterns with respect to the user's location. The only exceptions are the finance and traffic cards, that are split roughly 50-50 between home and office.

The left and middle plots in Figure 3 reveal strong temporal changes in card usage. However, they provide no details about temporal dynamics of engagement by the same user. In Figure 4 we investigate the distribution of temporal gaps between two consecutive clicked proactive impressions from the same user. On the

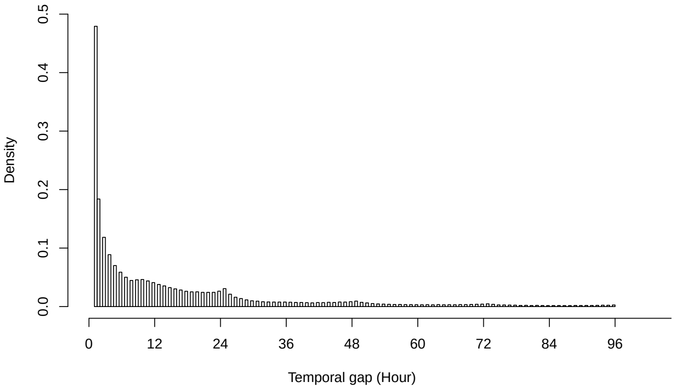


Figure 4: The temporal gap between the last click and the next click on cards. The x-axis is in hours and the y-axis shows the density of clicks that fall in hourly interval.

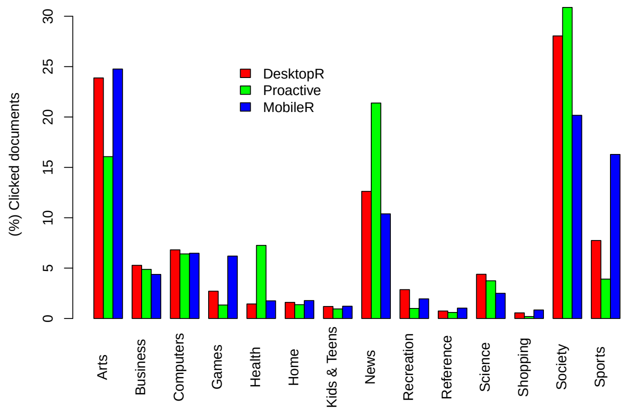


Figure 5: The distribution of ODP categories for documents clicked across different platforms.

x-axis we have the temporal gap (in hours), and on the y-axis we show the density of consecutive impressions that fall in that bucket. The results show that once the user clicks on a card, the next clicked impression is most likely to appear within an hour. It is also interesting to note the spikes around 24, 48, and 72 hours, suggesting that many users engage with the cards at the same time of the day.[3]

Overall, the results summarized in this section suggest that the dynamics of user interactions with the cards, follow similar patterns to those observed previously in reactive search logs.

## 5. CROSS-PLATFORM CLICKS

Thus far we have shown that the user engagement with the cards varies depending on the time and to a less extent based on location. Next, we investigate the topics of documents that users click on in their proactive impressions and compare them against their reactive history. It is important to note that not all the cards direct the user to a document. For instance, the traffic and finance cards direct the user to an app. Therefore, for this part of study, we limit ourselves to clicks from sports, news and places cards. We use the top level categories from the Open Directory Project (ODP) [3] to represent the document topics. To train topic classifiers, we followed the approach suggested by Bennett et al. [11] and used a 2008 crawl of ODP that was split into a 70%–30% train/validation set. We too discarded the `Regional` class as it is not topical which leaves us with 14 categories overall. For each topic, we train a logistic regression classifier with L2 regularization. We then run each clicked document in proactive and reactive impressions against all these classifiers and assign it to the topic that receives the highest score.

To distinguish between patterns that can be attributed to using different devices (Mobile vs. Desktop), rather than different platforms (reactive vs. proactive), we further divide the reactive user history based on the device on which the document was clicked. We respectively use DesktopR and MobileR to refer to documents that were clicked on desktop and mobile devices in reactive impressions. Figure 5 compares the topical categories of documents clicked on different platforms. It is evident that the overall distribution is different across platforms, although it is more similar between reactive impressions from mobile and desktop. There are certain categories such as `Health` and `News` that appear relatively more often among the documents clicked in proactive impressions. In contrast, categories such as `Art`, `Games` and `Sports` are relatively under-represented. The comparison between MobileR and DesktopR impressions confirm that users click on different types of documents on different devices. Similar argument can be made about the proactive-reactive differences although it should be noted that distribution of proactive clicks is also affected by what was presented to users in the cards.

Figure 5 illustrates the overall topical distribution of clicked documents aggregated over all users. However, it does not inform us about how the same user accesses information on different devices and platforms. Wang et al. [43] and Montanez et al. [33] studied the topics of documents that users click on cross-device. Here, we perform the same comparison cross-platform; that is, between proactive and reactive impressions. On each platform we discard all but the strongest ODP category associated with each user. For each pair $(M_p, M_q)$ of the three modalities, (Proactive, MobileR and DesktopR), we calculate the Normalized Point-wise Mutual Information (NPMI), to measure the association of a pair of ODP categories $(o_i, o_j)$ conditioned on the pair of modalities. The NPMI value is denoted as the output of function $N$, using the following formula:

$$N(o_i, o_j \mid M_p, M_q) = \frac{\log \frac{P(o_i, o_j \mid M_p, M_q)}{P(o_i \mid M_p, M_q) P(o_j \mid M_p, M_q)}}{-\log P(o_i, o_j \mid M_p, M_q)} \quad (1)$$

The value of $N$ ranges between $-1$ and 1, where 1 indicates that the two categories completely co-occur given the two modalities, $-1$ indicates that the two categories occur separately but not together, and 0 indicates that the two categories are independent. We visualize the cross-platform NPMI values for each pair of platforms and across all topics as heatmaps in Figure 6. The right diagonals of the three heatmaps have the darkest colors, indicating that, overall, the same category is clicked by the user across all pairs of modalities.

The consistency of the topics viewed cross-platform by the user suggests that the users' reactive history might provide useful signals for ranking their proactive recommendations. In the next sections, we put this assumption to the test.

[3] The data gets sparser and skewed towards 'heavy' users for longer periods. Therefore, we left the study of longer intervals (e.g. weekly/monthly) for future work.

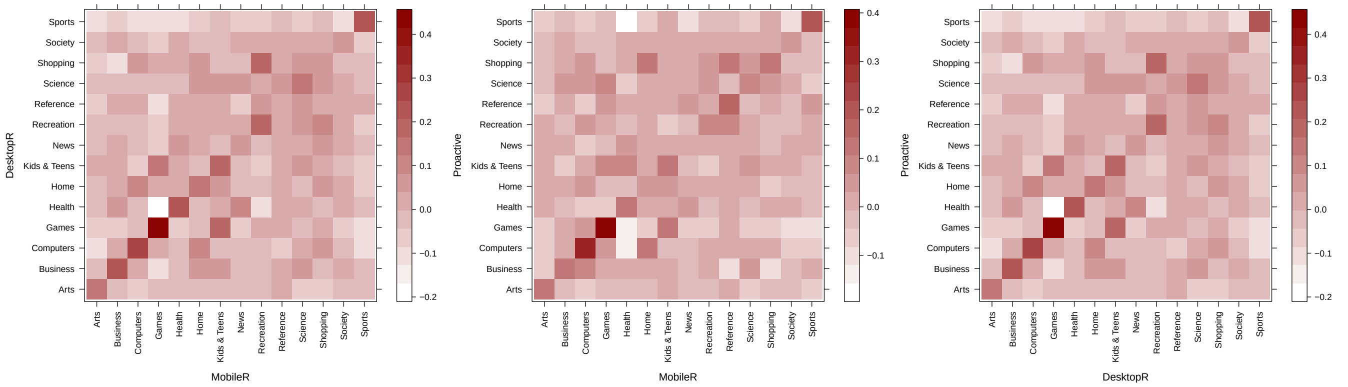


Figure 6: The ODP categories of documents clicked by the same user across different platforms. The colors in the heatmaps represent the normalized point-wise mutual information (NPMI) gains, where darker means higher.

Table 1: Offline labeling of proactive cards based on clicks (left). The ranking produced by Carré for the same impression (right). The reciprocal rank of each ranking is computed based on clicks and presented in the last row.

| Position | Baseline ($\pi$) | Clicked | Carré ($\pi^*$) |
|---|---|---|---|
| 1 | Weather | (✗) | Weather |
| 2 | Finance | (✗) | News |
| 3 | News | (✓) | Calendar |
| 4 | Calendar | (✗) | Finance |
| 5 | Places | (✗) | Places |
| RR ($E$) | 0.33 | | 0.5 |

# 6. RERANKING PROACTIVE CARDS

The results described so far suggest that the user's engagement with the cards is subject to various factors such as time, location, and position. They also reveal that the documents that are clicked by a user across proactive and reactive platforms tend to be consistent in terms of topicality. In this section, we use these insights to develop features that can be used to train a supervised model for reranking cards. We will refer to our trained model for card reranking as Carré.

*Formal definition.* Given a set of proactive cards $\mathcal{C}$, and the ranking produced over them by a default ranker $\pi(\mathcal{C})$, our task is to learn a reranking model $F$ (a.k.a. Carré) that reorders this list – if necessary – based on available features $\Theta$. The reranker is optimized towards a pre-defined metric $E$, so that its produced ranking $\pi^\star(\mathcal{C})$ outperforms the original order with respect to that metric. That is,

$$\pi^\star(\mathcal{C}) = F(\pi(\mathcal{C}), \Theta) \quad \textit{where,} \quad E(\pi^\star(\mathcal{C})) > E(\pi(\mathcal{C})) \qquad (2)$$

Therefore, to optimize Carré ($F$), the first question that needs to be addressed is the choice of optimization metric $E$ and labels.

*Pseudo-relevance labels.* Collecting context sensitive labels for training personalized rankers is not trivial. Different users in similar context (or even the same user in different contexts) may prefer different cards. The previous work in reactive search personalization addresses this problem by relying on clicks and inferring pseudo-relevance labels [12, 13, 38, 42]. Fox et al. [22] showed if a user dwells on a clicked document for longer than 30(s), that document is likely to be relevant. Inferring implicit labels for proactive cards is an open question. We apply the following three strategies in our experiments:

- **SAT-Clicks ($\mathcal{R}_c$):** We adopt the common click-based strategy for labeling cards. For each proactive impression in our logs, we consider clicked cards with $\geq$ 30(s) dwell time as relevant and others as non-relevant.

- **SAT-Views ($\mathcal{R}_v$):** The main problem with SAT-clicks is that there are many type of cards that do not require a click to satisfy the user's information need. For instance, although clicking on the weather card provides more details about the upcoming forecast, the current temperature is often what the user is looking for and it is always available on the card. This is analogous to the *good abandonment* phenomenon [30] in aggregated search, where the results of some verticals may not require a click to satisfy users. To remedy this issue, we rely on the viewport information to measure how long the user had the card visible on the screen during a proactive impression. Lagun et al. [29] demonstrated that viewport duration and gaze duration are highly correlated. While the authors did not specify a particular threshold for mapping viewport duration to relevance, previous work suggests that for short text segments – roughly the size of a paragraph – display time of $\geq$ 30(s) is a good indicator of relevance and is potentially more accurate than gaze for that purpose [16]. Therefore, we generate a second set of labels based on the viewport time. Cards with viewport duration $\geq$ 30(s) are considered as (pseudo) relevant and the others as non-relevant.[4] The red bars in Figure 2 depict the distributions of SAT-Views generated from viewport duration for cards presented at different positions. The trends are similar to what we observed for SAT-Clicks (blue bars) although the drop at position two is more substantial.

- **SAT-Hybrid ($\mathcal{R}_h$):** This approach combines the two options above and considers all cards with a SAT-Click or a SAT-view as (pseudo) relevant.

[4] Experiments using a tighter threshold ($\geq$ 15 seconds) produced similar results, and hence are excluded.

Table 2: The features used for reranking cards by Carré. The suits in brackets specify the experimental group that each feature belongs to. Note that the textual and topical features are only available to cards that recommend web documents and are left blank otherwise.

| Feature | Group | Description |
|---|---|---|
| ReactiveImpressionCount | (♣) | No. previous reactive search impressions by the user. |
| DesktiopImpressionCount | (♣) | No. previous reactive search impressions by the user issued from desktop. |
| MobileImpressionCount | (♣) | No. previous reactive search impressions by the user issued from a mobile device. |
| ReactiveDomainMatches | (♣) | No. domains (e.g. `cnn.com`) that appear in the card and have been clicked by the user in reactive logs. |
| PrevCardClicks | (♠) | Number of previous clicks by the user on this type of card. |
| ProactiveImpressionCount | (♠) | No. previous proactive impressions by the user. |
| ProactiveDomainMatches | (♠) | No. domains (e.g. `cnn.com`) that appear in the card and have been clicked by the user in proactive logs. |
| HuDist | (♥) | 2 features indicating the distance from the user's home, and office (in kilometers). |
| IsTimeOfDay | (♥) | 4 binary features indicating the time of day ∈{morning, afternoon, evening, night}. |
| IsDayOfWeek | (♥) | 7 binary features indicating the day of week ∈{Sat, Sun, $\cdots$, Fri}. |
| TemporalGap | (♥) | Time elapsed since the last proactive impression by the user (in hours). |
| Position | (■) | The position of card in the impression presented to the user. |
| CardCount | (■) | The number of cards presented to the user in the proactive impression. |
| IsCardType | (■) | 12 binary features generated for the most popular cards (e.g. IsSports). |
| IsOtherType | (■) | A binary feature that is set to one only if none of the IsCardType features is 1. |
| NumActiveDays | (■) | No. days with at least one impression in user's proactive and reactive history. |
| ImpressionNo | (■) | No. previous proactive impressions in the same session. |
| StaticRank | (■) | Computed based on number of SAT-Clicks and SAT-Views recorded on that type of card during the training period. |
| ReactiveTermMatch | (♦, ♣) | No. overlapping terms in URL-titles presented in the card and clicked URLs in reactive history (weighted by freq). |
| ProactiveTermMatch | (♦, ♠) | No. overlapping terms in URL-titles presented in the card and clicked URLs in proactive history (weighted by freq). |
| ReactiveCatMatch | (♦, ♣) | Similar to ReactiveTermMatch but computed based on the ODP categories assigned to URLs. |
| ProactiveCatMatch | (♦, ♠) | Similar to ProactiveTermMatch but computed based on the ODP categories assigned to URLs. |

While admittedly none of these approaches are ideal, we believe that they provide a reasonable benchmark for first step experimentation in this direction. We leave further investigation of the right labeling strategy as future work.

Given that than 2/3 of proactive impressions in our logs have only one card with a positive label (Hybrid-SAT = *true*), we decided to pick mean reciprocal rank (MRR) as our evaluation metric $E$. RR is defined as the reciprocal rank of the highest-ranked relevant card in an impression, and is considered as 0 in the absence of positive labels. MRR is computed by averaging RR over the test impressions. Our choice is in line with several previous works on search personalization that also used MRR as their evaluation metric [12, 38, 39, 42].

*Training.* To train Carré, we sampled proactive impressions from the logs of a commercial search engine as described in Section 3. Each impression consists of a unique and persistent user identifier that can be matched against reactive logs, a set of cards that were presented to the user, and the related click statistics and time-stamps. The relevance labels are assigned to the cards based on clicks and viewport duration as explained above. We then train a model that learns to rank the (pseudo) relevant cards on top based on available features and penalizes their demotion. Table 1 illustrates our labeling and evaluation process in a toy example. Here, the sampled impression contains 5 cards that were presented to a user. Among them, the user had a SAT-Click (or a SAT-View) on the news card which was shown at the third position (Hence, RR = $1/3 = 0.33$). Suppose that Carré has taken this impression along with the available features and reranked it as it is displayed on the right column of the table. Here, since the news card is placed higher at the second position we consider the ranking to be more effective than the original order (RR = $1/2 = 0.50$).

We chose LambdaMART [47] as our preference learning model. LambdaMART is a learning to rank approach based on gradient boosted decision trees and is an extension to LambdaRank [15]. It has been shown to be one of the most effective models for learning to rank [18], and has been a common choice among previous work on personalized ranking [12, 13, 38, 39, 42]. We performed a parameter sweep over the number of trees {20,100,500}, number of leaves per tree {2,4,8,$\cdots$,128}, the learning rate {0.02,0.05,0.1,0.4} and the minimum number of instances per leaf {1,10,50} over our validation datasets to tune the parameters. We split the proactive logs explained in Section 3 based on time-stamps into three sets for training, validation and testing. Impressions collected between 15 Nov 2014 – 21 Nov 2014 were used for training (291,039 in total), those collected between 22 Nov 2014 – 28 Nov 2014 (248,353 in total) were used for validation, and the remainder of impressions logged between 29 Nov 2014 – 7 Dec 2014 (272,289 in total) were used for testing and generating the reported results.

*Features.* Our experiments in the previous sections highlighted various factors that can affect the user engagement with the cards. Inspired by those observations we develop several features that can be used for card reranking in Carré. The comprehensive list of features is provided in Table 2, although in general they can be all summarized in five overlapping groups:

- Reactive history (♣): These are features that are generated based on the reactive search history of the user collected separately as described in Section 3.
- Proactive history (♠): These are features that are generated based on the last two weeks proactive history of the user.
- Lexical/Topical features (♦): These features are generated based on lexical and topical similarity between the URLs presented in the card, and documents clicked by the same user in previous proactive and reactive impressions. They are only available for cards that contain URL links and are set to zero elsewhere.
- Local/Temporal features (♥): These features are generated based on the time and the location of user when the proactive impression was presented.
- Constant features (■): Features that do not belong to any of the categories above (e.g. card position).

We can draw connection between many of the features used in our model to those deployed previously in the search personalization literature. For instance, the ProactiveCatMatch and ReactiveCatMatch features are similar to category-query features used by Bennett et al. [11]. The ReactiveDomainMatches and ProactiveDomainMatches features are related to the personal navigation models [41] but here we compute them at the domain level. Moreover, Position and ImpressionNo are both commonly used for reranking search results [11, 12, 38, 39, 42]

*Baselines.* We compare our results against two baselines. Our first baseline which we refer to as the *Default-ranker* is the original ranking of cards before ordering (middle column in Table 1). The Default-ranker is the production ranker of a commercial system (Cortana) that is already trained to produce the best ranking of cards at each impression. As our second baseline we use a static ranking of cards which produces the same ordering across all impressions. We count the number of times that each type of card receives SAT-Clicks or SAT-Views in our training period, and then rank the cards in all our testing impressions accordingly. We refer to this baseline as *Static-ranker* in our experiments. The effectiveness of Static-ranker highlights how well such non-personalized usage-based rankers can perform in the absence of other contextual signals.

## 7. EXPERIMENTS

We now focus on evaluating the effectiveness of our card reranking models. We begin our analysis by comparing Carré against our baselines. Given the proprietary nature of the Default-ranker, we do not disclose its absolute MRR, but instead compare the performance of other models against it in terms of relative changes in MRR. The results in Table 3 show that regardless of how the (pseudo) relevance labels are assigned, Carré outperforms the Default-ranker. The gains range between 6%-10% in terms of MRR and are all statistically significant according to the t-test ($p < 0.01$). The Static-ranker is significantly outperformed by the Default-ranker on all metrics ($p < 0.01$) which confirms the quality of the original rankings produced by the production system.

To verify if the gains and losses are distributed uniformly across users we decided to group users based on their *loyalty*. We ranked users in terms of the number of their proactive impressions in our logs. We then assigned the users in the top 15% and the bottom 15% of this ranking to separate groups that we respectively refer to as *Head* and *Tail*. Table 4 compares the performance of different approaches on Head and Tail user groups. Once again, the results are presented in terms of relative changes in MRR with respect to Default-ranker and all differences are statically significant except for one case (Carré, Tail users, $\mathcal{R}_v$). Carré consistently outperforms the Default-ranker on both sets of users. The gains are more substantial on the Head set across all metrics. This is expected as most of the proactive/reactive history (♣, ♠) and lexical/topical features (♦) are set to their default values for the Tail users due to lack of enough historical data. In addition, it can be noted that the gains are higher on the click-based evaluations ($\mathcal{R}_c$). Given that clicks can be considered as a more direct feedback from the user compared to the relatively more noisy labels inferred from viewport, perhaps it is not surprising to see the click-based models to be most effective. The performance of the Static-ranker is less affected by the user group, although the overall numbers follow similar trends.

Table 3: The effectiveness of card re-ranking models against the Default-ranker. The gains and losses are only reported in relative delta values (ΔMRR%) to respect the proprietary nature of the baseline ranker. The MRR values computed based on SAT-Clicks, SAT-Views and SAT-Hybrid label sets ($\mathcal{R}_c$, $\mathcal{R}_h$, $\mathcal{R}_h$) are grouped separately. All differences are statistically significance ($p < 0.01$) according to the pair t-test.

| | ΔMRR ($\mathcal{R}_c$) | ΔMRR ($\mathcal{R}_v$) | ΔMRR ($\mathcal{R}_h$) |
|---|---|---|---|
| Static-Ranker | ▼ -0.72% | ▼-13.24% | ▼-1.59% |
| Carré | ▲10.85% | ▲ 6.88% | ▲ 8.71% |

Table 4: The effectiveness of Carré against the baseline ranker for different groups of users (Head vs. Tail). The gains and losses are only reported in relative delta values (ΔMRR). Statistically significance gains and losses according to the pair t-test ($p < 0.01$) are respectively denoted by ▲ and ▼.

| | ΔMRR ($\mathcal{R}_c$) | ΔMRR ($\mathcal{R}_v$) | ΔMRR ($\mathcal{R}_h$) |
|---|---|---|---|
| *Head users* | | | |
| Static-Ranker | ▲ 2.12% | ▼ -2.82% | ▲ 1.52% |
| Carré | ▲19.69% | ▲16.35% | ▲17.69% |
| *Tail users* | | | |
| Static-Ranker | ▼-1.57% | ▼-8.86% | ▼-3.47% |
| Carré | ▲ 3.96% | 0.22% | ▲ 0.79% |

*Feature analysis.* We demonstrated that the rankings produced by Carré consistently outperform both baselines on all metrics and user groups. Next we investigate which feature groups contribute most to these gains.

We train separate rankers by excluding all features from each feature group and compare them with a Carré model trained over the complete feature set. The results are summarized in Table 5. As before, statistically significant differences (losses) according to the t-test ($p < 0.01$) are distinguished by ▼. Dropping the features from the constant group (■) hurts the performance more than any other subset regardless of the choice of labels. In particular, Position (original rank) is the most influential feature which is consistent with prior work on search personalization [12, 13, 38, 42]. The temporal/local (♥) features have the least impact on performance which is expected as none of them are card-specific. Features based on reactive history of the user (♣) are at least as effective as those generated based on the proactive history (♠). This is notable as it suggests that the previous search history of the users can be used for reranking cards, and is particularly useful for the first time users of proactive recommendations. Dropping all lexical/topical features (♦) computed over the reactive and proactive search history of users, substantially degrades the performance on all metrics. This confirms that the fact that the same user browses similar topics across platforms (Figure 6) can be used to improve their proactive recommendations.

## 8. DISCUSSION

The results presented so far, confirm the effectiveness of Carré for reranking proactive cards across all metrics. However, as we noted in Section 6, the click-based (pseudo) relevance labels may not be appropriate for evaluating all types of cards, and while the viewport-based measures partially address this problem, they are yet to be carefully tested on proactive scenarios. The threshold we picked was based on previous work on short segments of text [16], however different values might be discovered to be more suitable in proactive mobile settings.

Table 5: The effectiveness of card re-ranking models trained excluding the specified features against the Carré ranker trained with all features. The gains and losses are only reported in relative delta values (ΔMRR) against the Carré model trained with all features. Statistically significant differences ($p < 0.01$) according to the pair t-test are denoted by ▼.

| | ΔMRR ($\mathcal{R}_c$) | ΔMRR ($\mathcal{R}_v$) | ΔMRR ($\mathcal{R}_h$) |
|---|---|---|---|
| Carré (−♠) | ▼-1.04% | -0.04% | ▼-0.69% |
| Carré (−♣) | ▼-0.99% | -0.25% | ▼-0.94% |
| Carré (−♦) | ▼-7.15% | ▼ -5.71% | ▼-6.58% |
| Carré (−♥) | ▼-0.28% | 0.23% | ▼-0.13% |
| Carré (−■) | ▼-7.85% | ▼-10.06% | ▼-7.90% |

Table 6: The accuracy of news click prediction models trained excluding the specified features against the classifier trained with all features (bottom row). All differences are statistically significant according to the McNemar's Chi-Square ($\chi^2$) test.

| | Accuracy | ΔAccuracy |
|---|---|---|
| Carré (−♠) | 0.7980 | ▼-0.86% |
| Carré (−♣) | 0.7988 | ▼-0.77% |
| Carré (−♦) | 0.7932 | ▼-1.46% |
| Carré (−♥) | 0.7997 | ▼-0.65% |
| Carré (−■) | 0.7829 | ▼-2.74% |
| *The accuracy when trained with all features: 0.8050* | | |

To remedy this issue, we decided to conduct another experiment in which we exclusively focus on predicting clicks on news cards. Focusing solely on news would allow us to simplify the problem to a binary classification task. We also would not need to worry about the viewport-based metrics, as our news cards are clickable[5] and take the users directly to documents for which we know SAT-click signals can be used reliably to infer relevance [11–13, 22, 39, 46]. Therefore, we updated our experimental datasets described in Section 3 to form new sets that only include data for the news cards. The details of training/validation/testing split remain unchanged and the list of features stays identical. Instead of training a ranker based on boosted decision trees (LambdaMART [47]), we train a binary classifier based on gradient boosted decision trees [23]. The parameter sweep is performed over the same set of ranges that was specified in Section 6.

The motivation here is that if clicks on news cards can be predicted accurately, one can use this information to demote/promote their positions accordingly. The results in Table 6 show the click prediction accuracy of multiple Carré classifiers trained over different subsets of features. Accuracy is defined as the ratio of cards that were correctly classified ($\frac{\textit{True positives}+\textit{True negatives}}{\textit{All}}$) and the bottom row contains the accuracy of Carré when trained with all features. McNemar's Chi-Square ($\chi^2$) confirms that all differences are statistically significant ($p < 0.01$) against this full model that we consider as our baseline.[6]

Overall, the results are consistent with our previous experiments. Excluding the constant (■) and lexical/topical (♦) features degrades the accuracy more than any other feature group. The temporal/local (♥) features have the lowest impact while proactive (♠) and reactive (♣) features make similar contributions to accuracy. Using AUC as metric instead of accuracy led to similar conclusions. For instance, AUC dropped from 0.6909 to 0.6660 (▼-3.60%) in the absence of lexical/topical features (♦), and dropped to as low as 0.5987 (▼-13.34%) when constant features were excluded. We skip more details for brevity.

## 9. CONCLUSIONS

In this paper, we presented the first analysis on user interactions with proactive cards. We found several resemblances between the user interaction patterns with proactive cards and those recorded from reactive search logs. We demonstrated that similar to reactive search results, proactive card rankings also have a top-heavy click distribution. We also showed that similar to search queries, the card usage patterns are affected by temporal and local dynamics. For instance, there are certain cards (e.g. finance) that are likely to be clicked by the user in the morning, or at work. Furthermore, we investigated the topics of documents clicked by the user in different platforms and discovered that users are consistent across platforms in terms of the topics they interact with.

In the second half of our study, we used the insights obtained in the first half to develop features that can be used for reranking cards. We proposed a new approach for inferring the relevance of a presented card based on the user's click and viewport duration. We then deployed these implicit labels in a supervised framework for reranking cards. Our experimental results confirmed the effectiveness of our method (Carré) and showed that it consistently produces better results than the original ranker (and a static baseline). They also confirmed that features generated based on the reactive search history of users can be used to improve their proactive card recommendations.

There are several directions for future work. Inferring implicit relevance labels for proactive cards is still an open question. We tried to address that by adopting common techniques and thresholds based on previous studies on text, and also focusing exclusively on the news cards for some parts of this work. However, further investigations would be needed to better understand the user's interaction with the cards in order to derive realistic user models and metrics. While all our experiments were based on synthetic inferred labels, it would be important to assess our findings with real users in a user study or on live user traffic. In addition it would be useful to collect the proactive and reactive history of users over an extended and consistent range of dates to measure the impact of short-term and long-term history in each of these platforms. Last but not least, we demonstrated that the reactive history of users can be used to improve their proactive ranking; it would be interesting to investigate if the reverse holds true as well.

*Acknowledgments.* The authors are grateful to Peter Bailey, Sebastian Blohm, and Zeynab Raeesy for their thoughtful feedback on the first draft of this paper.

## References

[1] Google, Inside AdWords. http://bit.ly/1JrajCg/. Accessed: 2015-05-14.

[2] Cortana - Microsoft - USA. http://www.microsoft.com/en-us/mobile/campaign-cortana/. Accessed: 2015-05-14.

[3] DMOZ - the Open Directory Project. http://www.dmoz.org/. Accessed: 2015-05-14.

[4] Google now. https://www.google.com/landing/now/. Accessed: 2015-05-14.

[5] We ignore the potential *headline glance* interactions here.

[6] We used the t-test previously for measuring the statistical significance of differences in our ranking experiments. However, it is not applicable to this classification task.

[5] Apple - iOS 8 - Siri. https://www.apple.com/ios/siri/. Accessed: 2015-05-14.

[6] 2013 Strategic Imperative: Marketing as Applied Science | RKG Blog. http://bit.ly/1AnYMNZ, 2013. Accessed: 2015-05-14.

[7] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. SIGIR*, pages 3–10, Seattle, WA, 2006.

[8] J. Allan, B. Croft, A. Moffat, and M. Sanderson. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. *SIGIR Forum*, 46(1):2–32, May 2012.

[9] J. Arguello, J. Callan, and F. Diaz. Classification-based resource selection. In *Proc. CIKM*, Hong Kong, China, 2009.

[10] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Proc. SIGIR*, pages 321–328, Sheffield, UK, 2004.

[11] P. N. Bennett, K. Svore, and S. T. Dumais. Classification-enhanced ranking. In *Proc. WWW*, pages 111–120, Raleigh, NC, 2010.

[12] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *Proc. SIGIR*, pages 135–144, Beijing, China, 2011.

[13] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proc. SIGIR*, pages 185–194, Portland, OR, 2012.

[14] J. Budzik and K. Hammond. Watson: Anticipating and contextualizing information needs. In *Proc. ASIS*, volume 36, pages 727–740, 1999.

[15] C. Burges, R. Ragno, and Q. Le. Learning to rank with non-smooth cost functions. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, January 2007.

[16] G. Buscher, L. van Elst, and A. Dengel. Segment-level display time as implicit feedback: A comparison to eye tracking. In *Proc. SIGIR*, pages 67–74, 2009.

[17] L. D. Catledge and J. E. Pitkow. Characterizing browsing strategies in the world-wide web. *Comput. Netw. ISDN Syst.*, 27(6):1065–1073, Apr. 1995.

[18] O. Chapelle and Y. Chang. Yahoo! Learning to Rank Challenge Overview. In *Proceedings of the Yahoo! Learning to Rank Challenge*, pages 1–24, Haifa, Israel, 2011.

[19] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proc. WWW*, pages 1–10, Madrid, Spain, 2009.

[20] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *Proc. KDD*, pages 1082–1090, San Diego, CA, 2011.

[21] F. Diaz. Integration of news content into web results. In *Proc. WSDM*, pages 182–191, Barcelona, Spain, 2009.

[22] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst.*, 23(2), Apr. 2005.

[23] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[24] Q. Guo, H. Jin, D. Lagun, S. Yuan, and E. Agichtein. Mining touch interaction data on mobile devices to predict web search result relevance. In *Proc. SIGIR*, pages 153–162, Dublin, Ireland, 2013.

[25] M. Halvey, M. T. Keane, and B. Smyth. Time based patterns in mobile-internet surfing. In *Proc. CHI*, pages 31–34. ACM, 2006.

[26] K. Hofmann, B. Mitra, F. Radlinski, and M. Shokouhi. An eye-tracking study of user interactions with query auto completion. In *Proc. CIKM*, pages 549–558, Shanghai, China, 2014.

[27] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. *Inf. Process. Manage.*, 36(2):207–227, Jan. 2000.

[28] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2), Apr. 2007.

[29] D. Lagun, C.-H. Hsieh, D. Webster, and V. Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proc. SIGIR*, pages 113–122, Gold Coast, Queensland, Australia, 2014.

[30] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *Proc. SIGIR*, pages 43–50, 2009.

[31] H. Lieberman. Autonomous interface agents. In *Proc. CHI*, pages 67–74. ACM, 1997.

[32] Q. Mei and K. Church. Entropy of search logs: How hard is search? with personalization? with backoff? In *Proc. WSDM*, pages 45–54, Palo Alto, CA, 2008.

[33] G. D. Montanez, R. W. White, and X. Huang. Cross-device search. In *Proc. CIKM*, pages 1669–1678, Shanghai, China, 2014.

[34] A. K. Ponnuswami, K. Pattabiraman, Q. Wu, R. Gilad-Bachrach, and T. Kanungo. On composition of a federated web search result page: Using online users to provide pairwise preference for heterogeneous verticals. In *Proc. WSDM*, Hong Kong, China, 2011.

[35] F. Radlinski and T. Joachims. Query chains: Learning to rank from implicit feedback. In *Proc. KDD*, pages 239–248, Chicago, Illinois, USA, 2005.

[36] P. Resnick and H. R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, Mar. 1997.

[37] B. J. Rhodes and P. Maes. Just-in-time information retrieval agents. *IBM Syst. J.*, 39(3-4):685–704, July 2000.

[38] M. Shokouhi. Learning to personalize query auto-completion. In *Proc. SIGIR*, pages 103–112, Dublin, Ireland, 2013.

[39] M. Shokouhi, R. W. White, P. Bennett, and F. Radlinski. Fighting search engine amnesia: Reranking repeated results. In *Proc. SIGIR*, Dublin, Ireland, 2013.

[40] X. Su and T. M. Khoshgoftaar. A survey of collaborative filtering techniques. *Adv. in Artif. Intell.*, 2009:4:2–4:2, Jan. 2009.

[41] J. Teevan, D. J. Liebling, and G. Ravichandran Geetha. Understanding and predicting personal navigation. In *Proc. WSDM*, pages 85–94, Hong Kong, China, 2011.

[42] L. Wang, P. N. Bennett, and K. Collins-Thompson. Robust ranking models via risk-sensitive optimization. In *Proc. SIGIR*, Portland, OR, 2012.

[43] Y. Wang, X. Huang, and R. W. White. Characterizing and supporting cross-device search tasks. In *Proc. WSDM*, pages 707–716, Rome, Italy, 2013.

[44] I. Weber and C. Castillo. The demographics of web search. In *Proc. SIGIR*, pages 523–530, Geneva, Switzerland, 2010.

[45] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc. WWW*, pages 21–30, Banff, Alberta, Canada, 2007.

[46] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang. Enhancing personalized search by mining and modeling task behavior. In *Proc. WWW*, pages 1411–1420, Rio de Janeiro, Brazil, 2013.

[47] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Inf. Retr.*, 13(3):254–270, June 2010.

[48] S. Xu, H. Jiang, and F. C. M. Lau. Mining user dwell time for personalized web search re-ranking. In *Proc. IJCAI*, pages 2367–2372, Barcelona, Catalonia, Spain, 2011.