

Exploring and Exploiting User Search Behavior on Mobile and Tablet Devices to Improve Search Relevance

¹Yang Song, ¹Hao Ma, ²Hongning Wang, ¹Kuansan Wang

¹Microsoft Research,
One Microsoft Way,
Redmond, WA 98052, USA

²Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana IL, 61801 USA

ABSTRACT

In this paper, we present a log-based study on user search behavior comparisons on three different platforms: desktop, mobile and tablet. We use three-month search logs in 2012 from a commercial search engine for our study. Our objective is to better understand how and to what extent mobile and tablet searchers behave differently than desktop users. Our study spans a variety of aspects including query categorization, query length, search time distribution, search location distribution, user click patterns and so on. From our data set, we reveal that there are significant differences between user search patterns in these three platforms, and therefore use the same ranking system is not an optimal solution for all of them. Consequently, we propose a framework that leverages a set of domain-specific features, along with the training data from desktop search, to further improve the search relevance for mobile and tablet platforms. Experimental results demonstrate that by transferring knowledge from desktop search, search relevance on mobile and tablet can be greatly improved.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Internet search*; H.3.0 [Information Storage and Retrieval]: General—*Web search*

General Terms

Measurement, Experimentation

Keywords

mobile search, tablet search, user behavior analysis, search result ranking

1. INTRODUCTION

With the prevalence of smart phones and tablet PCs in the past few years, we have all witnessed an evolution in the search engine industry where the user search activities have shifted from desktop to mobile devices at an incredibly fast pace. According to a recent report [1], the year-on-year growth of mobile search volumes have more than doubled

from 2011 to 2012, while the search volumes on all platforms have merely increased by 11% over the year. Therefore, it is quite evident that understanding user search behavior on mobile devices has become more and more crucial to the IR community as well as the success of search engine companies. Particularly, what do users search? How do users formulate/reformulate queries? More importantly, what are the differences between desktop search and mobile search?

Google and Yahoo! released their statistics regarding mobile search usages [24, 14], in 2008 and 2009, respectively. The report from Google revealed that iPhone users bear lots of similarities with desktop users in terms of query length and query type distribution, while other mobile phones have shown different search patterns. On the other hand, Yahoo!'s findings disclosed that the US mobile query categorization is noticeably different from those of international mobile queries, where US users often issue longer and more complicated queries. At a high level, both reports have shown that users exhibit different search intent on mobile than desktop. For example, personal entertainment is the most popular category on mobile.

While much has been revealed from the aforementioned reports, we still believe that it is valuable to re-visit this problem with the latest data. As suggested in the Yahoo! report [24], the authors discovered that mobile search pattern is still evolving, in terms of query distribution they studied. One reason to support that is the query length: Yahoo! reported an average of 3.05 query length on mobile in 2008, which turns out to be 2.93 (for iPhone) and 2.44 (for mobile) in Google's 2009 report, indicating that users continue to change their search and reformulation behavior.

Moreover, at the time the above reports were written, tablet PCs were not so popular. However, with the debut of Apple's iPad in 2010, the usage of tablets for search has soared drastically in the past three years. It is reported that as of March 2012, over 30% of US users have one or more tablets [2]. Therefore, in this paper, we also analyze the behavior of tablet searchers in order to update the mobile search picture with this important missing piece.

Specifically, we make the following contributions:

- We collect search engine logs with over 1 million users for a period of three months (August 2012 to October 2012) from Bing desktop and mobile search engine, where we restrict our study to be English queries in the en-US search market. The mobile search logs include cell-phone users while desktop logs contain both desktop and tablet users.
- We perform a series of thorough and rigorous analysis

on mobile search behavior, in terms of time distribution, search locality, query categories, click patterns, browse patterns and so on, where we also compare with previous studies.

- Based on the results which revealed noticeable differences in mobile and tablet search patterns, we propose a novel knowledge transfer framework to train new rankers for mobile and tablet search results by leveraging a set of novel device-specific features, as well as incorporating the training labels from desktop data to improve the search relevance.

2. RELATED WORK

In this section, we review the literature of mobile search behavior in recent years. **Although there has been many related work** on mobile search [17, 5, 9, 19, 10, 11, 20, 18], we decided to primarily focus on the study after 2007 since the screen of smart phones have drastically changed after the appearance of iPhone in January 2007, which is one of the primary devices we analyze in this paper.

Kamvar and Baluja was among one of the earliest to report a large-scale study of mobile search statistics using over 1 million randomly sample queries from Google’s mobile log in early 2007 [13]. They explicitly compared the search behavior in 2007 to those of 2005, and found out that users tended to enter a query faster in 2007 due to the availability of high-end devices. They discovered that query length increased from 2.3 in 2005 to 2.6 in 2007. They also found out that, surprisingly, the click through rate, i.e., percentage of queries that had one or more clicks, had dramatically increased from less than 10% to over 50%. Their study also revealed more tail queries issued during 2007, as well as a larger portion of adult queries in mobile devices.

In 2008, Church et al. [8] reported a study on European mobile search logs, using 6 million queries from 260,000 search engine users over a period of 7 days. In their study, query length appeared to be similar for both desktop and mobile searchers, who also exhibited similar click patterns via focusing on top-ranked results. The query categories, or topics, however, appeared to be quite different, where adult queries consisted of over 60% of all queries, followed by email-related queries and personal entertainment queries. Their data also demonstrated a larger portion of navigational and transactional queries in mobile logs, consisted of 60.4% and 29.4%, respectively.

In [24], the authors studied Yahoo! mobile search logs during a period of 2 months in the second half of 2007, containing both US searchers and international searchers. They explicitly studied three user interfaces of mobile applications: Yahoo! oneSearch XHTML/WAP interface, Yahoo! Go for Mobile, and Yahoo! SMS Search. The authors have observed that personal entertainment was the most popular query category, after filtering out adult and spam queries. Comparatively, US searchers often issue longer queries with more words than international searchers, resulting more tail queries in the US search logs. At the same time, international searchers have larger diversity in terms of search intent, as indicated by the query topical distributions. Finally, the authors concluded that mobile search was still evolving, based on the inconsistency observed from a variety of studies.

The study in [14] by **Kamvar, Kellar, Patel and Xu** was among the first to make explicit comparison between iPhone users and other mobile users. They used the data from Google mobile logs **with** a period of 35 days in the sum-

mer of 2008, which contains over 100,000 queries issued by over 10,000 searchers. Their study revealed several interesting aspects. First, queries issued by iPhone users had similar length with desktop searchers (2.93), but significantly shorter for other mobile searchers (2.44). Second, iPhone searchers also exhibited similar query categories with desktop users, both of which were more diverse than mobile searchers. In terms of session statistics, the authors discovered that desktop users have the **most** queries per session, followed by iPhone then mobile users, indicating a possibility that the information needs were more diverse on desktop and iPhone, whereas mobile users were more likely to issue simple navigational queries with a more focused intent.

The authors in [15] studied user behavior in terms of abandoned queries in both mobile and PC search. The objective of the study was to approximate the prevalence of good abandonment, i.e., queries that lead to satisfied user information need without clicks. In three locales, US, JP and CN, the authors discovered that on mobile the portion of good abandonment is significantly higher than PC search. In particular, for the US search market, the highest rate of good abandoned queries were primarily from local, answer and stock search, which was different from the other two markets. The study suggested that query abandonment should not be treated as negative signal uniformly, instead both the locale and modality should be considered when categorizing abandoned queries into good and bad.

More recently, Teevan et.al [23] addressed the issue of local search on mobile devices, in which they conducted a survey on 929 mobile searchers. The authors claimed that the mobile local search experience differed a lot from desktop due to the limitation of the device, and therefore affected user search behavior for local-related information. Particularly, location (geographic features) and time (temporal aspects) played an important contextual rule in user’s search behavior. Users are more likely to search for locations that are close to their relative locations. Besides, users are more likely to initiate a local search during a specific time period.

Nevertheless, among all the aforementioned work we covered, none of them had studied the search behavior on tablet PCs such as iPad. In what follows, we will analyze the differences between mobile phone users and tablet users in a variety of aspects, as well as comparing to desktop searchers.

3. DEVICE AND DATA SET

Due to the availability of different types of mobile devices in terms of screen size and network capacity, it becomes difficult to study user behavior across different platforms. Thus we are unable to draw conclusions that are applicable to all types of devices. Consequently, we restrict our study by focusing on two devices that are most popular on the market, i.e., iPhone and iPad. The other reason is due to the consistent screen size of iPhone, which has not changed since its first generation¹, neither has iPad. Therefore, it is much more convenient and reliable to study the user search experience on these two devices.

In this paper, the data we use is from two different sources of Bing search logs. iPhone users who visit Bing are presented with Bing mobile search interface (<http://m.bing.com>), while iPad users are experiencing the same search interface

¹For iPhone 5, users see the same number of search results as other iPhone users but two more extra lines.

	Mobile (iPhone)	Tablet (iPad)	Desktop
Total Queries	9,732,938	8,423,111	13,928,038
Total Users	1,233,720	1,153,270	1,181,000

Table 1: The data sets used in this paper.The first row represents total query volume and the second row are number of unique users.

as desktop searchers (<http://www.bing.com>) but on a smaller screen. We distinguish iPhone and iPad users from other mobile users by filtering based on the request agent and the platform. We extract a sample of three months search logs in the United States search market, during August 2012 to October 2012 for iPhone, iPad and desktop users, respectively. We then sub-sample 1,000,000 users for each of the logs based on their unique user id string. Finally, we filter out non-English queries and consider only sessions that start from the Web vertical. Table 1 presents the overall statistics.

Note that from now on, “mobile” explicitly means “mobile phones” in our description, as to be distinguished from “tablet”. We will also be using terms “mobile” and “iPhone”, “tablet” and “iPad”, interchangeably.

4. USER BEHAVIOR ON MOBILE & TABLET

In this section, we analyze the differences between mobile and tablet searchers in a variety of aspects, including search time, location, query categorization, clicks and etc.

4.1 Query Distributions

In this section, we provide query statistics including query length, query categorization and etc for the three platforms.

4.1.1 Query Length

Table 2 summarizes the query length in terms of words and characters for mobile, tablet and desktop, respectively. We observe that mobile users in general issue longer queries than tablet and desktop users. The average number of words for mobile is 3.05, which is identical to Yahoo!’s report [24] and slightly larger than Google’s report (2.93) [14]. The length of tablet queries is shorter than mobile but longer than desktop, an average of 2.88 words and 18.02 characters per query.

The inconsistency of query length among various reports can be attributed to many reasons. e.g., the evolving of user typing behavior, the diversity of query intent on different platforms and so on. Among all these aspects, we believe that query auto-suggestion plays an important role. We discover that in our data set, query reformulation rate is almost identical on three platforms. However, mobile users are more likely to rely on auto-suggestion for query reformulation, perhaps due to the difficulty of typing [14], where often longer queries are suggested by the Bing search engine. On the other hand, since tablet users experience the same search interface as desktop users, it is hence not surprising to see the query length to be similar on these two platforms.

4.1.2 Query Categorization

To analyze the query categorization for each device, we use a multiple-class classifier which categorizes a query into over 80 different categories. Our classifier works slightly

	Mobile	Tablet	Desktop
Number of words	3.05	2.88	2.73
Number of characters	18.93	18.02	17.44

Table 2: Average Query Length.

Category	Mobile	Tablet	Desktop
Adult	23.5%	5.6%	5.0%
Autos	2.4%	2.7%	2.1%
Celebrity	8.3%	4.5%	3.0%
Commerce	8.6%	11.6%	7.7%
Finance	0.4%	1.0%	1.0%
Health	1.7%	2.3%	1.7%
Image	42.0%	25.8%	19.9%
Local	10.3%	11.5%	9.1%
Maps	0.1%	0.3%	0.3%
Movie	1.6%	0.9%	0.8%
Music	3.4%	2.3%	2.3%
Name	7.3%	4.8%	3.6%
Sports	5.7%	4.8%	3.8%
Navigational	15.4%	32.6%	36.9%

Table 3: Query Categorization Distribution.

different than previous ones [24, 14] in the sense that we use a different taxonomy than others, e.g., we do not have an “Entertainment” category but similar ones like “Celebrity” or “Game”. Additionally, our classifier also allows one query to be classified into more than one categories, e.g., query “Michael Jackson” is categorized into both “Celebrity” and “Name” categories.

Table 3 lists the top-14 categories and their corresponding query distributions. Different from previous reports [14], we see a significant difference in iPhone categories than desktop categories, while the difference between iPad and desktop is much less comparatively.

In general, mobile users are much more likely (23.5%) to issue adult-content queries than tablet and desktop users, which aligns with the fact found previously. Mobile users are also over two times more likely to search for celebrity. Also, 42% of the queries on mobile contain image-intent, where those queries trigger image answers on search result pages. Noticeably, desktop only consists of 19.9% image-intent queries. The most surprising finding for mobile queries, however, is the percentage of navigational queries, i.e., 15.4%, which is much less than iPad (32.6%) and desktop (36.9%). This discovery is interesting and also truth-telling. Our hypothesis is that, after digging into the data, for the iPhone platform which has a quite mature app market, developers have already released corresponding apps for those navigational queries such as the facebook app for query “facebook” and amazon app “amazon”. As a consequence, with those free and powerful apps in hand, iPhone users are more likely to directly use the apps for their tasks, instead of resorting to search engines to find the corresponding sites. Since navigational queries are in general shorter than informational queries, this further explains why iPhone users have longer query length as shown in Table 2.

On the other hand, iPad users exhibit different query category distributions than desktop in primarily two classes: local and commerce. First, our data confirms the findings

	Mobile	Tablet	Desktop
Mobile	/	58	47
Tablet	58	/	63
Desktop	47	63	/

Table 4: Top-100 Query Overlap on three platforms.

in [14] that for local-intent queries, only 1.2% more local queries were issued from iPhone users than desktop computers (1.7% in previous report). Surprisingly, iPad users have the largest percentage of local queries (11.5%) among all three, which could attribute to the fact that they use the map application more often than iPad and computer users [14]. Secondly, we observe that iPad users exhibit noticeably stronger shopping intent by searching for more commerce-related topics (11.6%) than the other two platforms.

Quantitatively, by normalizing the overall probability distribution to sum up to 1 and calculating the Kullback–Leibler (KL) divergence, we observe that desktop and mobile have the largest divergence in query distribution with a score of 0.31, followed by the score between tablet and mobile of 0.21. Tablet and desktop in overall exhibit quite similar query distributions with a KL score of 0.07.

Additionally, we list in Table 4 the number of overlapped queries for the three platforms, by comparing the top-100 most frequent queries from each of them. Comparatively, mobile and desktop share only 47 common queries, the least among all three pairs, while tablet and desktop have overlapping on 63 top queries. These statistics are consistent with the KL divergence score listed above.

4.2 Usage Time Distribution

So what time of the day do users use their mobile phones and tablets for search? Figure 1 illustrates the query volumes as a distribution of the time of the day. Not surprisingly, the majority of desktop search occurs during normal working hours from **8AM to 5PM**. Meanwhile, as we can see, mobile search volume continues to rise starting from 5AM till 10PM of the day, and then declines during midnights. On the other hand, tablet usage shows a fairly different pattern. The search volume of iPad maintains relatively low during normal business hours, i.e., from 7AM to 5PM. It then rises sharply from 6PM to 10PM, which peaks at 7PM with over 9% of the traffic of the day. It is interesting to see that the increase of iPad search activities during the night compensates for the steep decline of the desktop search volume at the same time period.

Furthermore, we are interested in the distribution of query categorization at different time of the day. To do so, we separate one day into four groups with each group of 6-hour duration. Table 5 lists the top-3 query categories for each platform during the day. We omitted the *night group* (1AM to 6 AM) here since the search volume is not significant to draw a distribution. Overall, mobile searchers exhibit much diversified information need at different time of the day. While showing image and navigational-related intent during the mornings, mobile searchers are more likely to issue local queries in the afternoons, adult and music queries during the evenings. For tablet users, local queries have been observed as one of the top-3 categories throughout the day, whereas commerce-related queries start to emerge in the afternoon and evening groups. In contrast, desktop searchers

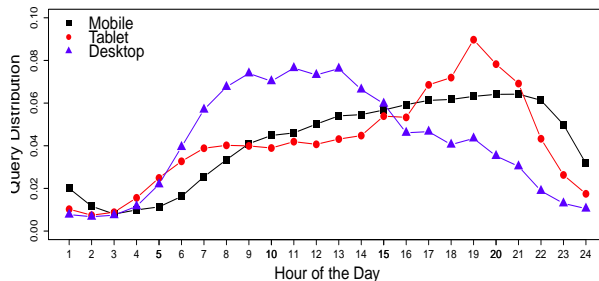


Figure 1: The time distribution of usage in terms of search volumes for three platforms.

	Mobile	Tablet	Desktop
Morning (7-12)	Image Navigational Name	Navigational Image Local	Navigational Image Local
Afternoon (13-18)	Local Celebrities Image/Name	Local Commerce Image	Navigational Local Name
Evening (19-24)	Adult Music Sports	Navigational Commerce Local	Navigational Local Commerce

Table 5: Break down of top query categorization by time of the day.

are always more interested in issuing navigational queries than the other two platform users, and exhibit a fairly stable information need throughout the day.

4.3 Location of Usage

Next, we analyze the location where search activities happen for mobile and tablets. As previous reports indicated, tablets are mostly used on couches and beds [3]. Therefore, we hypothesize that in terms of *mobility*, iPad users do not move as frequently as iPhone users, who can basically perform search at any location they go to during the day. To validate our assumption, we sub-select a sample of 2,000 users from the data sets in the Seattle area. For the mobile and tablet logs, we were able to extract the longitude and latitude of the location where the user query was issued. Each of the geo-location is mapped to its nearest city as well, e.g., “Redmond, WA”, “Bellevue, WA” etc.

Table 6 lists some statistics regarding location changes during the three-month logs we collected. Overall, mobile users searched from 4.52 different cities on average, while tablet users only traveled to an average of 1.79 cities. Note, it does not mean that on average a mobile user only appeared at 4.52 different places when performing search, since multiple geo-locations can be mapped to the same city name. Instead, the second row of the table shows how much distance (in miles) users traveled: on average mobile searchers were recorded to travel over 118 miles while tablet users only traveled for less than 50 miles. The percentage of users who never traveled, i.e., all search requests came from the same city, also shows significant difference between two types of devices. For mobile, less than 10% of users stay in one city, while the number is 3.7 times more for tablet users. Since

	Mobile	Tablet
Cities Visited	4.52	1.79
Distance Traveled (mi)	118.75	49.91
% users never traveled	9.54%	37.23%
% queries issued at home	43%	79%

Table 6: Statistics on location changes.

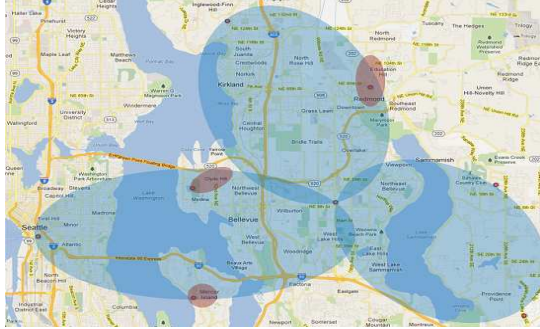


Figure 2: The locations where users performed search. Blue eclipses correspond to mobile searchers while red eclipses mean tablet searchers.

we do not know exactly where the “home” of a user locates, we simply treat the city that most queries were issued for that user as his/her home. By doing that, we observe that mobile users indeed issued queries at many more different locations than tablet users: 43% of queries issued at home versus 79%, which nearly doubled.

Figure 2 illustrates an example of three mobile users and three tablet users, respectively. We use eclipses to approximate the perimeters that cover the cities in which users traveled, where blue corresponds to mobile and red to tablet users. Clearly, the eclipses for mobile users cover much larger diameters than tablet users, who merely traveled to very nearby locations around their home.

4.4 Sessions and Clicks

In this section we analyze user session and click characteristics. We use the conventional definition of sessions in our analysis: a session contains user activities including query, query reformulation, URL clicks and so on. Sessions are grouped based on user IDs which are unique for each user. A session ends if there is no user activity for 30 minutes or longer. Each session is also assigned with a unique ID. The duration of a session is defined as the difference of timestamps between the last and first activity.

4.4.1 Session Duration and Engagement

Table 7 shows several basic statistics of sessions. Since a large portion of the sessions contain only abandoned queries, i.e., sessions with only one query but no clicks [15], we want to distinguish these sessions from others so we report numbers with and without these sessions. We first observe that mobile sessions contain the least number of queries (1.48), and therefore the shortest session duration of 7.62 minutes (8.25 filtered) among all three. iPad users, however, spent longer time than desktop and iPhone users with an overall 1.94 queries per session and 9.32 minutes (12.78 filtered) in each session. Despite longer session time, iPad users have

	Mobile	Tablet	Desktop
Number of queries in Session	1.48	1.94	1.89
Session Duration(min)	7.62	9.32	8.61
(Filtered) Session Duration(min)	8.25	12.78	10.22
Daily Number of Sessions	1.79	1.42	1.95

Table 7: Session statistics. Filter sessions are sessions without abandoned queries.

	Mobile	Tablet
SERP DwellTime (sec)	-20.54	+87.35
Avg Click Position	+ 0.54	-0.04
Answer CTR	+0.07	+0.02
Algo CTR	-0.20	-0.08
Click Entropy	+0.14	-0.05

Table 8: Click statistics are shown as relative numbers comparing to the desktop search numbers.

in general the least number of sessions in a day (1.42). This could be explained by the fact, as also shown in Figure 1, that most iPad search happened during night time between 7PM and 9PM so that the queries issued during that time are more likely to be in the same sessions. Comparatively, desktop users have more search sessions in a day than mobile and tablet which is expected since desktop still has the largest search volume.

4.4.2 Click Distributions

Next, we examine the differences in search result clicks. Note that our data is collected from Web vertical only. However, since Web vertical sometimes also shows image, video or local results, we group these non-algorithmic clicks to be *answer clicks* in our analysis.

In Table 8, we illustrate basic click statistics. Due to the sensitivity of the click data, we decide to report relative numbers here which use the number from desktop search as baseline comparison. Our first observation comes from SERP dwell time, where mobile searchers spent 20 less seconds examining the result page than desktop searchers. Meanwhile, tablet users exhibited much stronger interests in examining the SERP, 87 seconds more than desktop searchers. These numbers further confirm our results in Table 7 regarding the overall session duration time. In terms of click position, we observe that mobile searchers are more likely to click results that rank lower, an average increased position of 0.54. Tablet users, however, show no significant difference than desktop with similar click position (-0.04).

The click through rate (CTR) is an important metric used to measure search relevance. We can see from Table 8 that both mobile and tablet searchers are more lean to click on answers rather than algorithmic results (i.e., the ten blue-links). Specifically, mobile users have demonstrated 7% more CTR on answers but 20% less CTR on algo results, whereas tablet searchers are 2% more likely to click answers but 8% less willing to click algo results. Finally, the click entropy scores [22] indicate that on mobile, the clicks are more spread out with a higher entropy score, whereas on tablets clicks are often concentrated on top results, similar to that of desktop clicks.

Mobile	Tablet	Desktop
youtube.com	youtube.com	facebook.com
en.wikipedia.org	en.wikipedia.org	yahoo.com
answers.yahoo.com	amazon.com	en.wikipedia.org
ehow.com	msn.com	youtube.com
imdb.com	ebay.com	walmart.com
amazon.com	imdb.com	ebay.com
wiki.answers.com	ehow.com	amazon.com
chacha.com	facebook.com	mail.google.com
facebook.com	craigslist.org	aol.com
myspace.com	itunes.apple.com	craigslist.org

Table 9: Top-10 click domains.

4.4.3 Click Intents

The previous results include user clicks on both answers and algorithmic results. Since we see a significant difference on algo result clicks, we will focus on click patterns only on algorithmic results in this section.

We first list the top-10 most clicked URL domains in Table 9. It is interesting to see that the top-two clicked domains on desktop (i.e., facebook and yahoo), have both been replaced by youtube and wikipedia sites on mobile and tablet. The results again reflect our discovery in Table 3, but from the perspective of clicks, that mobile and tablet both have lower navigational-type queries and therefore users are less likely to click on those sites where iPhone and iPad apps are already available.

Next, we can also clearly observe that on tablet, shopping-related sites such as amazon, ebay and craigslist are ranked much higher than desktop and mobile. This is also consistent with Table 3 where we show that tablet users in general have higher percentage of commerce-related queries.

Perhaps the most interesting discovery in Table 9 is those highly-ranked *knowledge base* sites on tablet and mobile, especially on mobile. In particular, we observe that among the 10 most popular sites on mobile, 6 of them (Wikipedia, Yahoo! answer, ehow, etc.) are knowledge base sites. Consequently, the click distributions of traditional navigational queries have changed accordingly on mobile and tablet platforms. For example, for the query “louis vuitton”, desktop searchers clicked on the first result and went to its official site with over 0.7 CTR. However, this number dropped to 0.66 for iPad and 0.54 for iPhone searchers. In turn, iPad and iPhone users clicked more frequently on the Wikipedia page of *Louis Vuitton*, with CTR of 1.5% and 4.3%, respectively, even that the page is ranked at the bottom of the search results. Even for the query “facebook” that has the strongest navigation intent among all queries, mobile and tablet still possess 3.5% and 2.7% CTR on the Wikipedia page, respectively. Table 10 lists some example queries and their CTRs on knowledge base sites for mobile, tablet and desktop respectively. We have noticed that our findings concur with some recent studies as well [21, 7].

5. IMPROVE RANKING ON MOBILE AND TABLET

From the analysis we have performed in the previous section, we have clearly observed different user search behavior on both mobile (iPhone) and tablet (iPad) than desktop

CTR for Knowledge Base Sites on			
Query	Mobile	Tablet	Desktop
hotmail	2.5%	0.8%	0.3%
microsoft	16%	11%	2.0%
usa	33%	23%	14%
facebook	3.5%	2.7%	0.2%
louis vuitton	4.3%	1.5%	0.0%

Table 10: CTR for knowledge base sites such as Wikipedia, Yahoo! answers and etc. Mobile and tablet users are much more likely to click on those sites than desktop searchers.

users, in terms of query categorization, click intent, time and location of search and etc. Consequently, due to the diversity of search intent, mobile and tablet users have incurred a significantly lower click through rate (CTR) on algorithmic results as shown in Table 8, due to the factor of using a unified ranker on all three platforms.

To further improve the search relevance for mobile and tablet users, we propose to optimize the (algorithmic) search results by (1) incorporating new features that consider a variety of search aspects including time, location, intent and so on, (2) adopting the existing relevance labels from desktop search to train new rankers, as inspired by [4, 6].

5.1 New Features for Mobile and Tablet

Inspired by the analysis in previous section, two sets of features are derived in our framework. Specifically, query attributes features measure the characteristics of the query itself, across three devices and at different time of the day. On the other hand, URL relevance features estimate the importance of URLs given a particular query. Table 11 lists the features and the equations to calculate them.

5.1.1 Query Attributes Features

$\mathbf{q}\text{-prob}(\text{Query}|d)$ and $\mathbf{q}\text{-prob}(\text{Query}|t)$ measure the query frequency on a particular device d at time t of the day, where d is the one of the three devices we considered here: mobile, tablet and desktop. t is a numerical value indicating the time windows of the day, which is split into four groups (morning, afternoon, evening and night), similar as shown in Table 5.

$\mathbf{q}\text{-prob-cross}(\text{Query}|d)$ and $\mathbf{q}\text{-prob-cross}(\text{Query}|t)$, on the other hand, measure the cross-device and cross-time probability of a query. Comparatively, $\mathbf{q}\text{-prob}(\text{Query}|d)$ estimates how important the query is comparing to *all other queries issued on the same device*, whereas $\mathbf{q}\text{-prob-cross}(\text{Query}|d)$ judges how likely this query is issued *on that particular device rather than other two devices*. Likewise for t . These four features together demonstrate the overall importance of a query in the entire data set.

$\text{CTR}(\text{Query}|t, d)$ signals the search intent of users when they are interested in the query. The higher the CTR is, the more likely users are clicking on related URLs. The CTR is estimated by averaging over CTRs on all return URLs for the query during time t on device d in the data set.

$\text{Entropy}(\text{Query}|t, d)$ calculates the entropy of a query. This is also a signal to measure the popularity of the query during different time of the day, given a particular device d .

5.1.2 URL Relevance Features

$\text{KL}(\text{Class}(\text{Query}), \text{Class}(\text{U}))$ measures the topical closeness between the query and the URL. As we mentioned before, our classifier assigns each query (as well as a URL) into one or more of the 80 categories, which is essentially a probability distribution over all topics. The smaller the KL score is, the more likely the query and URL are related.

$\text{click-prob}(\text{U}|\text{Query}, t, d, \text{loc})$ considers the probability a URL gets clicked when a user issues the query at time t on device d and at a specific location loc . We specify the location parameter at two levels: *city* and *state*. To be concrete, we assign each city a unique ID and calculate $\text{click-prob}(\text{U}|\text{Query}, t, d, \text{city})$. Likewise for the state level. At these two different granularities, we measure the locality effect of the URL clicks.

Comparatively, $\text{loc-prob}(\text{U}|\text{loc})$ is a query-independent metric that calculates the overall locality effect of a URL. Similarly, $\text{Entropy}(\text{U}|\text{loc})$ also measures how likely the URL gets clicked at location loc . These two features are also parameterized with two different location levels: city and state. Likewise, we also include query-independent features for time and device, which have similar equations and therefore omitted from Table 11.

Since we have discovered that mobile and tablet users are more likely to click on knowledge base sites, we propose a feature to take that into consideration. Specifically, $\text{wiki-prob}(\text{U})$ calculates the probability of a sites to be knowledge base, according to the frequency the site is clicked. We maintain a list of over 30 knowledge base sites including Wikipedia, Freebase, Yahoo! answers, eHow and etc.

5.2 Improve Ranking via Knowledge Transfer from Desktop Search

With the new features derived in the previous section, we are ready to train a ranking model to improve the relevance on mobile and tablet. One way to achieve this goal is to leverage the learning-to-rank framework [16] by collecting judgement labels for individual query-URL pairs to form a training set, and use the domain-specific features in Table 11 to train a ranker. However, this approach seems suboptimal due to the expensive cost of acquiring human labels. In particular, it is very labor-intensive and cost-ineffective to gather labels on mobile and tablet platforms, especially in our scenario where each query-URL pair can generate multiple labels according to different time, location and device. Consequently, the cost could go exponential to the number of query-URL pairs and make this approach unable to scale.

On the other hand, human judgement labels for desktop search are available in abundance on many benchmark data sets, e.g., TREC, LETOR [16] and etc. These data sets, along with a rich set of textual features such like BM25, term frequency and etc, facilitate the work of training rankers for desktop search using different machine learning methods.

Therefore, our objective is to leverage the labels and content-features from desktop search, combining with a few labels from mobile and tablet as well as their domain-specific features, to train new rankers for the two new domains. Our framework is greatly inspired by the work in [4] and [6]. The general idea is to simultaneously use the source (desktop) and target (mobile and tablet) domain training data during the learning process, where the training data from the target domain is difficult to acquire while the data from source domain is available in abundance. These two domains, however, share certain common features, and therefore we can

learn a low-dimensional representation so that the features from both domains can be projected to. In [4], the authors have shown that this problem can be formulated as a 1-norm regularization problem that provides a sparse representation for multiple domains. Furthermore, in [6], the authors have proven that for learning-to-rank, the same problem can be transformed into an optimization framework which can be solved by using Ranking SVM [12], after certain transformation of the training data.

Formally, we are given three sets of training data \mathbf{D}_d , \mathbf{D}_m and \mathbf{D}_t , which correspond to desktop, mobile and tablet respectively. These data sets share the same feature space $\mathbf{w} \in \mathbb{R}^d$, which can be broken down into two parts. The first part contains k features $[\mathbf{w}_1, \dots, \mathbf{w}_k]$ that are common features available on all domains (e.g., BM25, document length and etc). The second half of features $[\mathbf{w}_{k+1}, \dots, \mathbf{w}_d]$ are domain-specific features which are only available on mobile and tablet domains. The learning objective is to minimize the pair-wise loss on all three domains. Specifically, for each domain, given a set of training data $\mathbf{D} = \{\mathbf{x}_i\}_1^n$, we form a set of pair-wise preferences $\mathbf{S} = \{(\mathbf{x}_{i1}, \mathbf{x}_{i2})\}_1^n$, where each pair indicates a preference relationship of $\mathbf{x}_i \succ \mathbf{x}_j$, which can be determined, for example, using human labels. Using Ranking SVM as a learning framework, we can assume the learning function to be linear, e.g., $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{s} \rangle$, where $\mathbf{s}_i = \mathbf{x}_{i1} - \mathbf{x}_{i2}$ is a new training sample by subtracting the feature values of \mathbf{x}_{i2} from \mathbf{x}_{i1} . The label y_i of \mathbf{s}_i is 1 if $\mathbf{x}_i \succ \mathbf{x}_j$ and -1 otherwise. This way we form a set of new training data $\mathbf{S}' = \{\mathbf{s}_i, y_i\}_1^n$ solvable by Ranking SVM.

Algorithm 1 sketches the learning process of how to minimize the pair-wise loss for three domains. Our algorithm is similar to the CLRank algorithm in [6]. The major difference is that we apply the learning to three domains instead of two which was the case in [6]. The general idea is to find a lower-dimensional representation of the three feature vectors, by performing SVD on \mathbf{D} , which represents the covariance matrix of the model weights, or how many common features these three domains share. The training instances are then transformed into this low dimension and trained using Ranking SVM. After training, the original feature weights are updated by transforming back the weights into its original dimension. The matrix \mathbf{D} is also updated with the new \mathbf{W} . This algorithm runs in iterations and stops when some criteria are met, e.g., the covariance matrix \mathbf{D} is no longer showing significant change.

6. EXPERIMENTS ON RANKING

In this section, we conduct rigorous experiments to assess the performance of ranking mobile and tablet algorithmic results using domain-specific features and the CLRank algorithm. The data sets used for ranking is the same as the ones we perform user behavioral analysis, as shown in Table 1. To gather the desktop training data, we ask human assessors to manually label each query-URL pair with 5-point Likert scale: Perfect (5), Excellent (4), Good (3), Fair (2) and Bad (1). Each query-URL pair is given to three human assessors and we apply majority vote to get its final label. Overall, we randomly select 3,500 query-URL pairs from the 1 million queries used in our study for judgement.

On the other hand, as mentioned above, for mobile and tablet, it is difficult to collect human labels because each query-URL pair can have different ratings depending on the time, location, device and etc. As a result, we resort to user

Type	Feature
Query Attributes	$\mathbf{q}\text{-prob}(\text{Query} \text{d}) = \frac{\text{cnt}(\text{Query} \text{d}) + \lambda_{qd}}{\sum_q \text{cnt}(\text{Query} \text{d}) + \lambda_d}$
	$\mathbf{q}\text{-prob-cross}(\text{Query} \text{d}) = \frac{\text{cnt}(\text{Query} \text{d}) + \lambda_{qdc}}{\sum_d \text{cnt}(\text{Query} \text{d}) + \lambda_{cd}}$
	$\mathbf{q}\text{-prob}(\text{Query} \text{t}) = \frac{\text{cnt}(\text{Query} \text{t}) + \lambda_{qt}}{\sum_q \text{cnt}(\text{Query} \text{t}) + \lambda_t}$
	$\mathbf{q}\text{-prob-cross}(\text{Query} \text{t}) = \frac{\text{cnt}(\text{Query} \text{t}) + \lambda_{qtc}}{\sum_t \text{cnt}(\text{Query} \text{t}) + \lambda_{ct}}$
	$\text{CTR}(\text{Query} \text{t}, \text{d}) = \text{avg}[\text{CTR}(U \text{Query}, \text{t}, \text{d})]$
	$\text{Entropy}(\text{Query} \text{t}, \text{d}) = -\mathbf{q}\text{-prob}(\text{Query} \text{t}, \text{d}) * \log \mathbf{q}\text{-prob}(\text{Query} \text{t}, \text{d})$
URL Relevance	$\text{KL}(\text{Class}(\text{Query}), \text{Class}(U)) = \sum_c \log \left(\frac{P_c(\text{Query})}{P_c(U)} \right) P_c(\text{Query})$
	$\text{click-prob}(U \text{Query}, \text{t}, \text{d}, \text{loc}) = \frac{\text{cnt}(U \text{Query}, \text{t}, \text{d}, \text{loc}) + \lambda_{qtdl}}{\sum_{q'} \sum_{t'} \sum_{d'} \sum_{l'} \text{cnt}(U \text{Query}, \text{t}, \text{d}, \text{loc}) + \lambda_{qtdl'}}$
	$\text{loc-prob}(U \text{loc}) = \frac{\text{cnt}(U \text{loc})}{\sum_{l'} \text{cnt}(U \text{loc})}$
	$\text{Entropy}(U \text{loc}) = -\text{loc-prob}(U \text{loc}) * \log \text{loc-prob}(U \text{loc})$
	$\text{wiki-prob}(U) = \frac{\text{cnt}(U) * \mathbb{1}(\text{IsWiki}(U))}{\sum_{u \in \text{List}(\text{wiki})} \text{cnt}(u)}$

Table 11: List of query-attribute and URL-relevance features used for ranking. All the λ 's are smoothing parameters that are estimated from the data set.

Algorithm 1 The CLRank algorithm

- 1: **Input:** Three training set converted to Ranking SVM format
 $\mathbf{S}^d = \{\mathbf{s}_i^d, \mathbf{y}_i^d\}_1^{N_d}$, $\mathbf{S}^m = \{\mathbf{s}_i^m, \mathbf{y}_i^m\}_1^{N_m}$, $\mathbf{S}^t = \{\mathbf{s}_i^t, \mathbf{y}_i^t\}_1^{N_t}$,
Parameter γ for Ranking SVM
- 2: **OUTPUT:** ranking models on original features $\mathbf{W} = [\mathbf{w}_d, \mathbf{w}_m, \mathbf{w}_t]$
- 3: Initialize covariance matrix $\mathbf{D} = \frac{\mathbf{I}_{d \times d}}{d}$;
- 4: **while** not converge **do**
- 5: **for** $m = 1$ to M
- 6: Do SVD on \mathbf{D} , so that $\mathbf{D} = \mathbf{P}^T \mathbf{\Sigma} \mathbf{P}$;
- 7: Multiply all feature vectors in \mathbf{S}^d , \mathbf{S}^m and \mathbf{S}^t with $\mathbf{\Sigma}^{\frac{1}{2}} \mathbf{P}$, get three new training data sets $\mathbf{S}^{d'}$, $\mathbf{S}^{m'}$, and $\mathbf{S}^{t'}$
- 8: Run Ranking SVM on these data sets to get feature weights \mathbf{u}_d , \mathbf{u}_m and \mathbf{u}_t ;
- 9: Transform $\mathbf{w}_d = \mathbf{P}^T \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{u}_d$, $\mathbf{w}_m = \mathbf{P}^T \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{u}_m$, $\mathbf{w}_t = \mathbf{P}^T \mathbf{\Sigma}^{\frac{1}{2}} \mathbf{u}_t$;
- 10: Set $\mathbf{D} = \frac{(\mathbf{W}\mathbf{W}^T)^{1/2}}{\text{trace}(\mathbf{W}\mathbf{W}^T)^{1/2}}$
- 11: **end while**

clicks as *pseudo labels* for mobile and tablet. We count all clicks for each query-URL pair which is parameterized by time, location and device. For each query, we assign the same 5-point labels to URLs according to the descending order of the click counts. Overall, we collect 5,000 such query-URL pairs for mobile and tablet, respectively – a total of 10,000 training examples.

For each training instance, we generate 400 content-based features such like BM25, document length [16], along with the 20 domain-specific features proposed in Table 11.

6.1 Baseline Comparisons and Metrics

To be more convincing, we compare with several baseline methods in order to show the superior of our proposal.

Baseline 1: the default ranking model of mobile and tablet – the same ranker as the desktop search. Here we do not modify anything but just report the default score.

Baseline 2: content-based features plus new domain-specific features. i.e., we train Ranking SVM models for mobile and tablet respectively, using the new 5,000 training instances described in the previous section.

Baseline 3: knowledge transfer without new features.

Train CLRank algorithm on three domains with the original 400 content-based features. This model is similar to our final ranker except that it does not leverage the domain-specific features proposed in Table 11.

Note that for both baseline 1 and baseline 2, two separate rankers need to be trained respectively for mobile and tablet. While for baseline 3, as well as our final ranker, only one optimized model will be outputted, i.e., the \mathbf{W} matrix on line 2 of Algorithm 1, which contains feature weights for all three domains.

For evaluation, we employ two classic metrics: MAP@K and NDCG@K. MAP calculates the mean of average precision scores on all queries in the test set. NDCG score, on the other hand, takes both the 5-scale relevance score and the position of the relevant documents into consideration. In our experiments, we report results for both $K = 1$ and 3.

6.2 Experimental Results

To report experimental results that are statistically meaningful, we randomly separate the 10,000 labeled data of mobile and tablet into two parts for training and test at 1:1 ratio, where the test set is withheld only for evaluation. We repeat this process 20 times and report the average performance. To determine the optimal value of the only parameter γ in the CLRank algorithm, we perform a 5-fold cross validation on the training set, and find out $\gamma = 0.15$ to be the optimal value.

Table 12 compares the overall performance of the four methods on mobile and tablet, in terms of MAP and NDCG scores respectively. In general, we see that both baseline 2 and baseline 3 make noticeable improvement over the default baseline 1. In comparison, baseline 3, which applies the knowledge transfer framework (CLRank) without new features, slightly outperforms baseline 2 which only uses the new features to train new rankers. As mentioned previously, baseline 3 leverages the CLRank model that jointly optimizes the rankers for all three domains, instead of training separate rankers for each domain as used by baseline 2. This comparative result indicates the potential superiority of using existing labels from other domains to enhance the current ranking system of the target domain.

On the other hand, when combining the domain-specific

	Mobile				Tablet			
	MAP@1	MAP@3	NDCG@1	NDCG@3	MAP@1	MAP@3	NDCG@1	NDCG@3
Baseline 1	0.3725	0.3981	0.2988	0.3584	0.2693	0.2986	0.2381	0.2949
Baseline 2	0.3746	0.4082*	0.3074	0.3782*	0.2711	0.3001	0.2579*	0.3183*
Baseline 3	0.3843	0.4123	0.3076	0.3799*	0.2854*	0.3129*	0.2894*	0.3281*
Our Method	0.4226**	0.4526**	0.3412**	0.3985**	0.2973**	0.3285**	0.3189**	0.3498**

Table 12: Overall performance of three baseline methods and our framework in MAP and NDCG. Our method outperforms all baseline methods, where * indicates p -value < 0.05 and ** means p -value < 0.01 .

features with the labels from desktop training data, we have observed significant performance improvement of our method comparing with all the baselines, with statistical significance level at p -value < 0.01 for all the metrics. Overall, our framework improves around 5% for both MAP and NDCG on mobile ranking baseline 1, whereas for tablet, the improvement is less (3%) but still quite significant comparing to other baselines.

In previous experiments, we limit the use of desktop training data to be 3,500. It would be helpful to analyze the performance change when that part of data becomes more/less. Consequently, we run a series of experiments by using only a certain portion of the desktop data for knowledge transfer, ranging from 500 to 3,500 instances.

Figure 3 illustrates the MAP and NDCG changes in terms of the training data size. Note that neither baseline 1 nor baseline 2 leverages training data from desktop, their performance is therefore not affected as demonstrated by the horizontal lines. Overall, we observe that for both baseline 3 and our method, more desktop data indeed helps improving the performance. More specifically, for the mobile domain, we see a dramatic increase of MAP and NDCG scores when the data increases from 500 to 1,000. The performance is then stabilized after 2,000 instances and only minor improvement can be observed beyond that. Comparatively, we notice that for the tablet domain, the performance increase is almost linear to the number of desktop training data, where it shows no sign of stopping even when all 3,500 training data has been utilized. Therefore, the tablet domain may potentially benefit more if we can provide more than 3,500 labeled desktop data in this scenario.

Next, we illustrate the performance improvement within different query categories. In Figure 4, we show the improvement in terms of the MAP@3 scores by our method over baseline 1. Among all 14 categories, our algorithm improves mostly on navigational, local and map queries. As we discussed before, for navigational queries, mobile and tablet users tend to click more on the knowledge base sites. Therefore, by leveraging the domain-specific features to improve the ranking of these sites, we successfully increase the MAP score for those navigational queries. Comparatively, we also observe that tablet has less MAP@3 improvement. In particular, queries in adult, movie and name categories benefit little from our algorithm. It could be the reason that these queries are more informational, where user intents are more diversified and therefore more difficult to optimize.

Finally, we also break down the metrics based on the time of the day as discussed in Section 3, as shown in Table 13, which indicates the relative improvement of our CLRank algorithm over the default baseline method. We see that in general, mobile gains the most improvement during after-

	Mobile		Tablet	
	MAP@3	NDCG@3	MAP@3	NDCG@3
Morning	0.082	0.032	0.028	0.053
Afternoon	0.126	0.065	0.021	0.032
Evening	0.097	0.043	0.055	0.074
Night	0.031	0.027	0.015	0.036

Table 13: MAP and NDCG improvement based on the time of day. The numbers mean the absolute difference between our algorithm and baseline 1.

noons and evenings, whereas tablet has the biggest jump in evenings. Since the majority of the search traffic is from afternoons (for mobile) and evenings (for mobile and tablet) as illustrated in Figure 1, we can clearly see the benefit of using our algorithm over the existing systems.

7. DISCUSSIONS AND CONCLUSIONS

The objective of our query analysis study was to reveal the user behavior difference between mobile, tablet and desktop searchers by using the latest log data from a commercial search engine. More importantly, our research aimed at filling the gap between mobile and desktop search by taking tablet user behavior into consideration, which we found missing in the previous studies of mobile search [24, 14]. We provided quantitative statistics on a variety of aspects for mobile and tablet search which were later used to guide the improvement of search result ranking on them. Our study on the three-month Bing mobile and desktop logs disclosed various points that were different from previous studies. The following is a few key discoveries from our study:

- The query length on mobile continues to change. Our study showed an average of 3.05 words per query for mobile and 2.88 for tablet, both of which are longer than desktop. Comparing to the numbers from Yahoo! (3.05) and Google (2.93), this number seems to change all the time. Therefore, we think that the usage patterns are still evolving.

- The distribution of query categories was different between mobile, tablet and desktop in certain categories. Specifically, tablet users were more likely to issue commerce and local queries, while mobile users issued more adult, celebrity and image queries. One important finding in our study was that both mobile and tablet users issued significantly less navigational queries than desktop users, due to the wide availability of mobile apps on these two platforms.

- The distribution of usage time was also different on three platforms. While desktop users performed search mostly during working hours (8AM to 5PM), mobile and tablet usages peaked during evenings (6PM to 10PM). We also

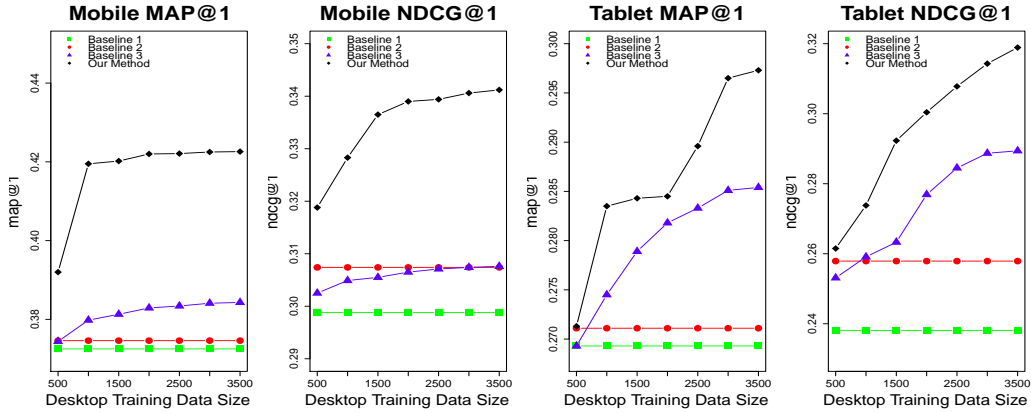


Figure 3: MAP@1 and NDCG@1 change in terms of the desktop training data size, which is used in the knowledge transfer algorithm (CLRank) to boost the performance of mobile and tablet.

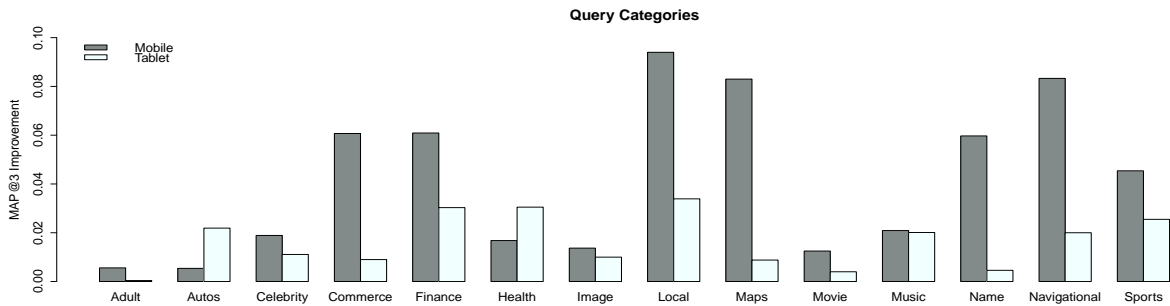


Figure 4: MAP@3 improvement by our algorithm over baseline 1 in 14 query categories.

revealed that during different time of the day, mobile and tablet users had more diverse search intent than desktop, where the latter seldom changed throughout the day.

- The location of usage was quite different between mobile and tablet. Overall, mobile users tended to travel a lot more than tablet users and issued queries at a variety of locations. For tablet users, 79% of queries were issued at home, whereas only 43% mobile users issued queries at their home locations.

- Interestingly, mobile and tablet users tended to click more on knowledge base sites like Wikipedia, even for very top navigational queries such as “facebook” and “hotmail”. This discovery can help us better understand user click intent which may eventually lead us to re-train the query classifier for mobile and tablet.

- Merely using traditional content-based features to rank search results on mobile and tablet can lead to suboptimal performance. We observed a significantly lower CTR on these two platforms while using the default ranker from desktop. We therefore proposed a set of domain-specific features to address the limitation of features from desktop. By leveraging a knowledge transfer algorithm (CLRank) that used training data from all three domains simultaneously, we eventually saw a significant performance improvement on mobile and tablet. This study revealed that (1) domain-specific features were important for mobile and tablet relevance. Even with only 20 new features, we have witnessed a 5% and 3% relevance improvement for mobile and tablet, re-

spectively, and (2) human labels from desktop can be leveraged to improve the rankers for other domains as well. A joint optimization on three rankers work better than optimizing rankers for different domains individually, especially when these domains share some common features.

Overall, we have observed that tablet users have distinguished themselves from desktop and mobile users with quite different user behavior and intent. It is therefore suggested that when performing user behavior analysis or designing relevance algorithms, tablet should be treated as a separate device rather than merge it with either desktop or mobile.

There is still a lot left to be done in the future. Our study covered two most widely used mobile and tablet devices (iPhone and iPad) on the market. However, with the choice of smart devices becomes more and diversified nowadays, it is important to take user behavior on all different devices into consideration to get a complete picture to draw unbiased conclusions. In the future, we plan to extend our work to cover more smart devices such as Android tablets, Microsoft Surface, iPad Mini and so on. A comparison between these tablet users may yield different outcomes than what have been observed in this paper. On the other hand, since we have observed that more and more users tend to click on answer results besides the algorithmic results, we will also be interested in extending our knowledge transfer algorithms beyond algorithmic results, by also adopting answers and ads into the framework.

8. REFERENCES

- [1] Mobile search volume more than doubles year-on-year. <http://www.smartinsights.com/search-engine-optimisation-seo/seo-analytics/mobile-search-volume-more-than-doubles-year-on-year/>.
- [2] Survey: 31 percent of u.s. internet users own tablets. <http://www.pcmag.com/article2/0,2817,2405972,00.asp>.
- [3] What you need to know about targeting ipad and tablet searchers. <http://searchengineland.com/what-you-need-to-know-about-targeting-ipad-tablet-searchers-109685>.
- [4] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*. MIT Press, 2007.
- [5] R. Baeza-yates, G. Dupret, and J. Velasco. A study of mobile search queries in japan. In *In Query Log Analysis: Social and Technological Challenges, WWW 2007*, 2007.
- [6] D. Chen, Y. Xiong, J. Yan, G.-R. Xue, G. Wang, and Z. Chen. Knowledge transfer for cross domain learning to rank. *Inf. Retr.*, 13(3):236–253, June 2010.
- [7] K. Church and N. Oliver. Understanding mobile web and mobile search use in today’s dynamic mobile landscape. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI ’11*, pages 67–76, New York, NY, USA, 2011. ACM.
- [8] K. Church, B. Smyth, K. Bradley, and P. Cotter. A large scale study of european mobile search behaviour. In *MobileHCI 2008*, pages 13–22, New York, NY, USA, 2008.
- [9] K. Church, B. Smyth, P. Cotter, and K. Bradley. Mobile information access: A study of emerging search behavior on the mobile internet. *ACM Trans. Web*, 1(1), May 2007.
- [10] Y. Cui and V. Roto. How people use the web on mobile devices. In *Proceedings of the 17th international conference on World Wide Web, WWW ’08*, pages 905–914, New York, NY, USA, 2008. ACM.
- [11] R. Hinman, M. Spasojevic, and P. Isomursu. They call it surfing for a reason: identifying mobile internet needs through pc internet deprivation. In *CHI ’08 Extended Abstracts on Human Factors in Computing Systems, CHI EA ’08*, pages 2195–2208, New York, NY, USA, 2008. ACM.
- [12] T. Joachims. Optimizing search engines using clickthrough data. In *KDD 2002*, pages 133–142, 2002.
- [13] M. Kamvar and S. Baluja. Deciphering trends in mobile search. *Computer*, 40(8):58–62, Aug. 2007.
- [14] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices. In *WWW 2009*, pages 801–810, New York, NY, USA, 2009.
- [15] J. Li, S. Huffman, and A. Tokuda. Good abandonment in mobile and pc internet search. In *SIGIR 2009*, pages 43–50, New York, NY, USA, 2009.
- [16] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, Mar. 2009.
- [17] Y. Lv, D. Lymberopoulos, and Q. Wu. An exploration of ranking heuristics in mobile local search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’12*, pages 295–304, New York, NY, USA, 2012. ACM.
- [18] H. Mueller, J. L. Gove, and J. S. Webb. Understanding tablet use: A multi-method exploration. In *Proceedings of the 14th Conference on Human-Computer Interaction with Mobile Devices and Services (Mobile HCI 2012)*, 2012.
- [19] H. Müller, J. Gove, and J. Webb. Understanding tablet use: a multi-method exploration. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services, MobileHCI ’12*, pages 1–10, New York, NY, USA, 2012. ACM.
- [20] S. Nylander, T. Lundquist, and A. Brännström. At home and with computer access: why and where people use cell phones to access the internet. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’09*, pages 1639–1642, New York, NY, USA, 2009. ACM.
- [21] C. A. Taylor, O. Anicello, S. Somohano, N. Samuels, L. Whitaker, and J. A. Ramey. A framework for understanding mobile internet motivations and behaviors. In *CHI ’08 Extended Abstracts on Human Factors in Computing Systems, CHI EA ’08*, pages 2679–2684, New York, NY, USA, 2008. ACM.
- [22] J. Teevan, S. T. Dumais, and D. J. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *SIGIR 2008*, pages 163–170, 2008.
- [23] J. Teevan, A. Karlson, S. Amini, A. J. B. Brush, and J. Krumm. Understanding the importance of location, time, and people in mobile local search behavior. In *MobileHCI 2011*, pages 77–80, New York, NY, USA, 2011.
- [24] J. Yi, F. Maghoul, and J. Pedersen. Deciphering mobile search patterns: a study of yahoo! mobile search queries. In *WWW 2008*, pages 257–266, New York, NY, USA, 2008.