Evaluating and Predicting User Engagement Change with Degraded Search Relevance

Yang Song
Microsoft Research
One Microsoft Way
Redmond, WA
yangsong@microsoft.com

Xiaolin Shi
Microsoft Bing
One Microsoft Way
Redmond, WA
xishi@microsoft.com

Xin Fu*
LinkedIn Corporation
2029 Stierlin Court
Mountain View, CA
xin.fu.2007@gmail.com

ABSTRACT

User engagement in search refers to the frequency for users (re-)using the search engine to accomplish their tasks. Among factors that affected users' visit frequency, relevance of search results is believed to play a pivotal role. While multiple work in the past has demonstrated the correlation between search success and user engagement based on longitudinal analysis, we examine this problem from a different perspective in this work. Specifically, we carefully designed a large-scale controlled experiment on users of a large commercial Web search engine, in which users were separated into control and treatment groups, where users in treatment group were presented with search results which are deliberate degraded in relevance. We studied users' responses to the relevance degradation through tracking several behavioral metrics (such as query per user, click per session) over an extended period of time both during and following the experiment. By quantifying the relationship between user engagement and search relevance, we observe significant differences between user's short-term search behavior and long-term engagement change. By leveraging some of the key findings from the experiment, we developed a machine learning model to predict the long term impact of relevance degradation on user engagement. Overall, our model achieves over 67% of accuracy in predicting user engagement drop. Besides, our model is also capable of predicting engagement change for low-frequency users with very few user signals. We believe that insights from this study can be leveraged by search engine companies to detect and intervene search relevance degradation and to prevent long term user engagement drop.

Categories and Subject Descriptors

G.3 [**Probability and Statistics/Experimental Design**]: controlled experiments, randomized experiments, A/B testing; H.4.m [**Information Systems**]: Miscellaneous

General Terms

Measurement, Design, Experimentation, Prediction

*Work done at Microsoft Bing.

WWW 2013, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2035-1/13/05.

Keywords

search quality, search relevance, user engagement, longitudinal analysis

1. INTRODUCTION

As Web search engines have become a necessity in our daily lives, the *relevance* of search engine results has undoubtedly become the deterministic factor for search engines like Google, Bing, Yahoo! and etc. to declare their successes and compete for query market share. For the past decade, researchers and practitioners have never stopped working towards improving search engines' relevance, by leveraging state-of-the-art methods from communities like machine learning, data mining, natural language processing and so on. These restless efforts have brought tremendous success to search engine companies with noticeable advances in search result relevance. According to a recent report by Experian[1], the *search success rate* for Yahoo!, Bing and Google were 81%, 80% and 66%, respectively.

Although it seems to be common sense that the better a search engine's relevance is, the more likely that users will engage with it (i.e., come back to search more often), it has come to our attention that few research effort has been spent to study the relationship between search relevance and user engagement. Until very recently, Hu et al.[6] proposed to characterize the relationship between *search success* and search engine reuse by measuring the correlation between changes in search satisfaction ratio and the rate of return. A positive correlation was identified between these two variables, which truly indicates that search success can lead to higher user engagement rate.

In this paper, we propose to study search relevance and user engagement from a ranking perspective, and in a more controlled environment. Despite the existence of many factors that can potentially influence search relevance, such as snippet quality, domain bias [7], the quality of the ranking algorithm is by far the most dominant component that determines a search engine's relevance score. We therefore isolate ranking from all other factors to better understand the change of user engagement from a longitudinal perspective.

On the other hand, our study tries to look at the correlations between these two variables from a different perspective than previous work [6]: how user engagement changes when the ranking algorithm suddenly becomes *worse than*

[1]http://www.experian.com/hitwise/press-release-experian-hitwise-reports-google-share-of-searche.html

before? While most existing research aims at improving the ranking algorithm, and most user engagement analysis is conducted in the environment where users succeeded their search objective, we, however, believe that *search failure* can potentially lead to engagement changes which is more complicated to understand and therefore should not be ignored. The momentum for this study is quite straightforward: search engine companies nowadays make changes to their ranking algorithms on regular basis. Before releasing a new ranking algorithm, it is common sense to first test it on a small portion of user basis, which is widely known as randomized experiments or A/B tests [10]. Apparently, not all changes can lead to an improved ranking algorithm in practice but may possibly hurt user experience and turn users away. It is therefore essential to understand the user behaviors under this scenario and make proper adjustments, e.g., performing early-termination of a bad ranking experiment, promptly.

Specifically, this paper makes the following contributions: We present a longitudinal study on a large amount of users from the logs of a widely-used commercial Web search engine. These users are enrolled in a carefully designed experiment in which the changes were *only* made to the ranking algorithm such that the ranking results look worse than before, while all other elements on the search result page stay the same. We then perform deep analysis on user engagement changes by studying the user behavior data at session-level, user-level as well as query-level to fully understand the root causes of user behavior changes.

Given the user engagement numbers over time, we propose a machine learning model to predict engagement changes, i.e., whether a user will come back more/less often in the future. To be concrete, we leverage a set of features such as the average length of user queries, the portion of queries that has no clicks and so on, and use these features to train an SVM model to make binary predictions for each user's *weekly* engagement change.

The rest of the paper is organized as follows: Section 2 discusses Related Work; Section 3 presents the details of our controlled experiments and the data collected; Section 4 performs deep analysis on the data; Section 5 proposes the model to predict user behavior changes; we conclude our paper and discuss potential future work in Section 6.

2. RELATED WORK

Our study is based on a large-scale controlled experiment of a commercial search engine. In the industry of online services, designing online experiments to test the impact of changes of products or services with large amounts of real users has been an extremely important problem [19]. Among many approaches, online controlled experiments have been widely adopted, as this type of experiments have the advantages of having best scientific design for establishing a causal relationship between changes and their influence on user-observable behavior and providing the first-hard feedback from large volumes of online users directly [10].

There has been extensive study on online user behavior that is related to information seeking and navigation. Most of such study uses recorded user search data [8], which can provide us with rich signals about different aspects of user behavior. For example, by studying clickthrough data, we are able to evaluate and monitor user satisfaction toward the relevance of a search engine [16, 20]. We can also have a good estimation on users intent [8] or how much they are frustrated with their search [5]. Previous research has also found that, by tracking user search behavior over time, such behavior could be well modeled and predicted [3, 15]. Moreover, some longitudinal study on user search behavior suggests that there are two classes of users: navigators and explorers [21]. In this work, we show that user behavior in the virtual world of information search, similar to behavior in many other real-world social and ecology systems [18], is also highly adaptive with regard to the change of this information system.

The main focus of user behavior in our study is user engagement in search, which directly ties to the market share of a search engine. It is believed that in many other forms of products and brands, customer satisfaction has a strong influence on the loyalty and market share [1, 14]. However, only very recently researchers started investigation on the relationship between user engagement and their satisfaction toward the service provided by search engines [6]. One significant difference between the research of user engagement and loyalty in the use of search engines and other products lies in that, it is a complicated problem of defining the usage of a search engine. This is because the frequency of issuing queries is not equal to the frequency of accomplishing tasks in information search, as users may issue multiple queries to accomplish one task [13, 11]. Moreover, we should be aware that, in real business, the long-term user behavior do not always align with short-term behavior, and this problem is particularly prominent in user engagement in search [10]. Thus, unlike [6], which only considers user usage on the query level, we focus more on the task or session level in terms of user engagement in this paper.

3. EXPERIMENT DESIGN AND DATA COLLECTION

In this work, we conducted an A/B testing user experiment on a widely-used commercial Web search engine in the US search market, from Jan 2011 to March 2011 for a total of 47 days. Two randomized buckets of users, of approximately equal size, were chosen as the *control* and *treatment* groups, whereas for the control group, the ranking algorithm remains the same as before. For the treatment group, we deliberately released an inferior ranking algorithm which has shown to have worse relevance scores in terms of NDCG scores [9] – approximately 3-point NDCG loss. To be more concrete, we used an old ranker with a less sophisticated machine-learning model and a different set of features. Almost 100% of queries were affected, i.e., showing different or re-ordered top-10 results between control and treatment. Figure 1 shows an example. For the query "yahoo email", we can clearly observe the differences between control and treatment results. For control, the ideal ranking is preserved where the #1 result is indeed what users look for. However, for users in the treatment group, they suffered from a totally weird list of ranked results: the official yahoo homepage was ranked first, followed by an ehow page and a wikipedia page, both of which are somewhat irrelevant. The fourth result looks like the desired page at the first glance, but unfortunately it turns out to be misleading too. Notice that although here we say that 100% of queries have different top-10 results, what we really mean is that *at least* one result in top-10 are different. Therefore, many of the queries,

especially top navigational ones, still show the same top-3 or top-5 results in both control and treatment. In the next section, we shall examine the impact of *truly* degraded queries to users in more details.

Overall, a total of 2.2 million users were enrolled in this experiment. After the finish of the experiment, we collected user session data from the logs. By our definition [6], a user session is a sequence of user search activities, including user queries, clicks on the search results and so on. A session ends when the user becomes inactive for 30 minutes or more. Users are anonymized by given a randomly generated user id for each user.

Let's define a metric named percentage delta, which is used throughout this paper to measure the difference of any metric between treatment and control:

$$\%\Delta(f) = \frac{f(Treatment) - f(Control)}{f(Control)} \times 100\% \quad (1)$$

Here f can be any arbitrary metric, e.g., session per user, click through rate and so on (details in next session). In our study, we notice a significant difference between the number of active days of users in treatment (4.500 days) vs. control (4.515 days) (whose % Δ is -0.335% with p-value 0.004), where a user is defined as active if he/she issued one or more queries on that day. However, there is no significant difference between the range of active days of users in treatment and control (both are around 11 days on average), which is defined as the date difference between user's first active day and last active day. For example, two users u_1 and u_2 are active on days $\{1, 2, 3, 4, 15\}$ and $\{1, 15\}$ of the experiment, respectively. Both of them have 15 as their range of active days, while u_1 has 5 active days but u_2 has only 2. These facts reveal that, the quality change of search engine has effect on the users' frequency of usage; however, we don't see there is a higher rate of abandonment when the search quality gets worse.

4. USER ENGAGEMENT ANALYSIS

We focus on three aspects of user analysis in this section: (1) treatment/control comparison in terms of several key metrics, (2) affected user analysis, i.e., how user engagement changes before and after users were exposed to bad relevance results, and (3) degree of affect analysis: quantifying engagement ratio changes across user groups characterized by the number of bad search results experienced.

4.1 Overall Engagement Changes Over Time

If we consider the search engine as an ecosystem, it is expected to see that users change their behavior accordingly once this ecosystem, i.e. the quality of the search results, changes. In this part, we investigate the overall user search behavior change with regard to the change of the search environment with a deliberate setback. In generally, search related user behavior can be classified into three categories:

a. Types of queries users issue. This type of behavior is performed before a user seeing the search result of the current query. Queries can be classified according to different criteria, such as navigational queries vs. informational queries, head queries vs. tail queries and short queries vs. long queries [2, 12].

b. User's satisfaction toward the search results. This type of behavior shows users' reaction toward the search result of the current query. Behaviors such as how quick the users click the search results and how soon they issue the next queries are strong signals showing users' satisfaction.

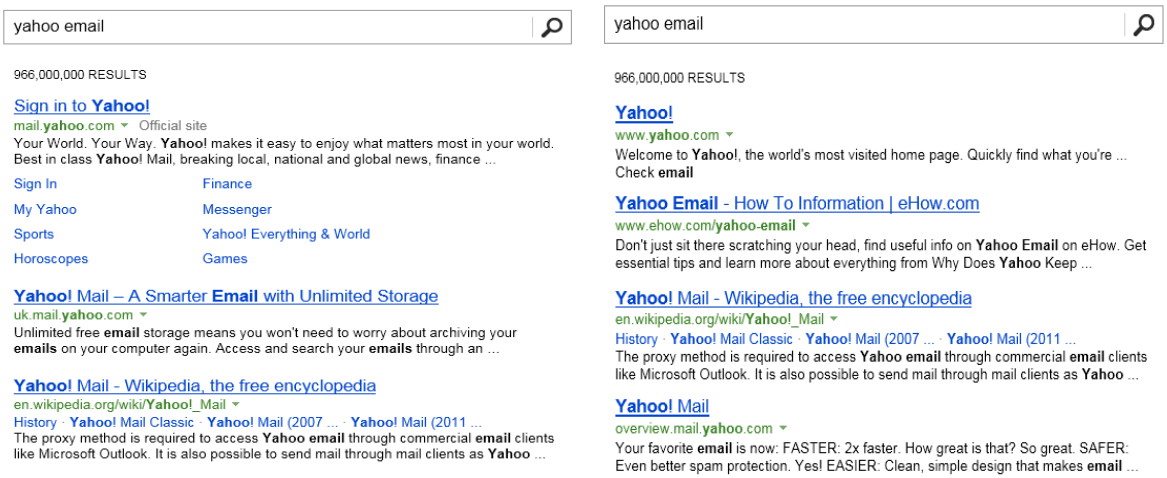
c. User's engagement in search. This type of behavior indicates the usage frequency and how frequently users using or reusing the search engine in order to accomplish their search tasks. From the following analysis, we will see that the short-term and long-term engagements are very different.

By studying the three types of user search behaviors, we aim to answer the following two questions: 1, what are the temporal patterns of user search behavior when the search quality changes? 2, how do different search behaviors correlate with each other? Especially, how do other behaviors correlate with the long-term user engagement?

Therefore, we focus on examining the following key metrics in our study (the type in square bracket indicates the type of user behavior mentioned above):

- **[c] Average daily sessions per user (S/U)**: $\frac{\sum_u S(u)}{|u|}$, where $S(u)$ indicates user u's daily session number and $|u|$ is the total number of users on that day.
- **[b] Average unique queries per session (UQ/S)**: $\frac{\sum_s UQ(s)}{|s|}$, where $UQ(s)$ represents the number of unique queries within session s, and $|s|$ the total number of sessions on that day.
- **[b] Average session length per user (SL/U)**: the total number of queries within a session, averaged over each user.
- **[a] Percentage of navigational queries per user (%-Nav-Q/U)**: there exists many methods for this type of classification [2, 12]. We propose a simple method by looking at click positions: if over n% of all clicks for a query is concentrated on top-3 ranked URLs, this query is considered to be navigational. Otherwise it is treated as informational. Here we empirically set n to be 80.
- **[a] Average query length per user (QL/U)**: the query length measures the number of words in a user query.
- **[b] Average query success rate per user (Q-Success/U**: a user query is said to be successful if the user clicks one or more results and stays at any of them for more than 30 seconds [6].
- **[b] Average query Click Through Rate (CTR)**: the CTR for a query is 1 if there is one or more clicks, otherwise 0.
- **[b] Average query interval per user (QI/U)**: the average time difference between two consecutive user queries within a user session.

While S/U indicates the user engagement ratio, the remaining metrics measure the user satisfaction as well as their effort spent to complete user's search task, and therefore can be categorized as *relevance* metrics.



(a) Control (original ranking)

(b) Treatment (worse result)

Figure 1: Example query of "yahoo email" and the ranking differences between control and treatment.

We first examine the user engagement change by looking at S/U and UQ/S. Rather than comparing treatment and control groups on a daily basis, we propose to use *accumulated numbers* to better capture the metric changes from a longitudinal perspective. For example, to compare S/U on the K's day since the starting of the experiment, we aggregate user sessions from day 1 to day K for each user, and calculate the difference between treatment and control using unpaired (two sample) t-test.

The accumulated results of these two metrics are shown in Figure 2. With the treatment group having a worse relevance algorithm, we initially expect to see a noticeable drop of of S/U in treatment in the first few days, resulting a dip in t-stat that may reach significance level (t-stat $\leq$ -1.96) very quickly. However, in contrast with common sense, we see that S/U in treatment actually went up quite a bit for the first three days, and then gradually decreased. The t-stat for S/U did not reach significance level until almost two weeks after the experiment. We also observe diminishing returns after two weeks for the rest of days.

On the other hand, UQ/S has shown significance from the very first day of the experiment, with t-stat > 5 for the first few days. We have observed that UQ/S raised over 2% for treatment comparing to control (3.66 vs. 3.57) on day 1, and continue to stay at 1% for the first two weeks. This shows a clear sign that users in treatment group have spent more effort to complete their search tasks than users in control.

Next, we show the (non-accumulative) temporal change of user behavior related to their satisfaction with the search results and engagement in Figure 3. We measure the difference between treatment and control using eq.(1). In Figure 3(a), we see that the average query success rate drops significantly as soon as the search quality degrades, which means that user satisfaction toward their search results is an immediate indication of the change of search quality. Figure 3(b) shows the change of average query intervals. The time interval is also strongly correlated with user satisfaction due to the fact that if users are not satisfied with their current search results, they are very likely to reformulate the original queries within a very short period of time. Similarly, Figure 3(c) shows that after the search results get worse,

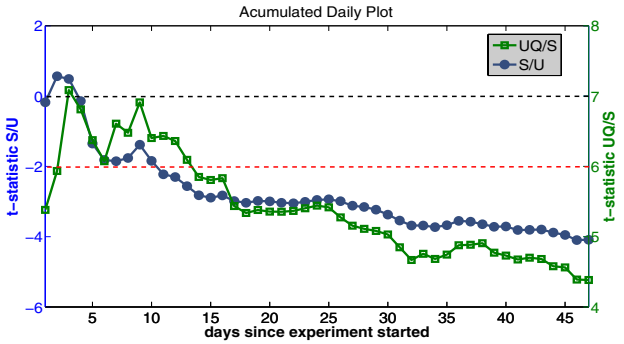


Figure 2: Accumulated daily plots for S/U and UQ/S. T-stats are shown on the left and right y-axis respectively.

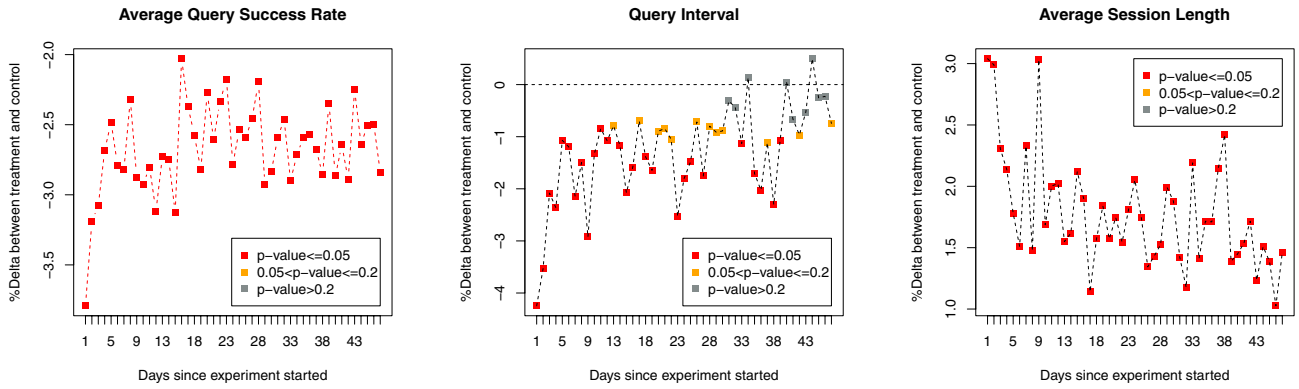
the average session length increases. This again tells us that after the search quality gets worse, there are more queries within a session as it requires more efforts from the users to accomplish their search tasks.

Therefore, with user engagement metric S/U going up then down (↑ then ↓), relevance metric UQ/S going up (↑), and average session length going up a little bit (↑), what is really going on underneath these metric changes? There can be a few possible reasons: (1) users indeed come back less frequently, but issue the same type of queries day after day. Due to the deteriorated relevance, users need to formulate more frequently to complete same tasks as before; or (2) users still come back at the same rate as before, but however do not *trust* the search engine any more, and hence give easier tasks to it, e.g., by issuing more navigational queries. For more difficult tasks, users simply find other workarounds for example by switching to another search engine with better relevance results.

To find the exact answer for that, we perform a more fine-grained analysis on *affected* users in the next section.

4.2 Affected User Analysis

What we have observed in the previous section is the s-



(a) (b) (c)

Figure 3: Daily change of behavior related to user satisfaction toward their search results and user engagement with search.

tatistics from overall users in both groups, which gives us some hints in terms of which direction those metrics developed. However, in this experiment we performed (as well as many other relevance-improvement experiments), though we claim that almost 100% of queries had different search results, only a small fraction of queries are *practically* involved, where their top-3 search results got re-ordered. Since most of the time, users are indeed only examine and click top-3 search results. As a result, many users were not truly exposed to results with bad relevance at all and their engagement changes are quite unnoticeable. Therefore, we only aim at studying behavioral changes for the group of affected users in this section.

We define the set of affected queries retrospectively. After the experiment was finished, we collected a set of queries that had been issued by users from both treatment and control groups. Within them, we isolated queries where the rankings of the results were different in treatment and control. We also filtered out queries with very low frequencies to make sure the difference was not caused by server instability. For each of these queries, we gathered top-10 results for treatment and control respectively. We then asked human assessors to evaluate the relevance of each result with respect to that query, in a 5-level Likert scale: Perfect (5), Excellent (4), Good (3), Fair (2) and Bad (1). We employed normalized discounted cumulative gain (NDCG) [9] to assess the relevance of a ranking result, where higher NDCG scores (between 0 and 1) indicate better relevance. With that, the set of affected queries is defined to have at least k difference ($k \in (0, 1)$) between control and treatment, with control having higher score. The value of k balances the trade-off between the size of the query set and the discriminative power of the set, where a high value of k results in a smaller set but queries with larger NDCG differences. Note that since we are only interested in truly impacted users, we use NDCG@3 as the metric here.

We then define the set of affected users. A user is said to be affected if he/she issued at least one of the queries in the affected query set. To better interpret user behaviorial changes before and after the user was affected, we select users who were only affected *after* the third week of the experiment. We expect those users to behave similarly as

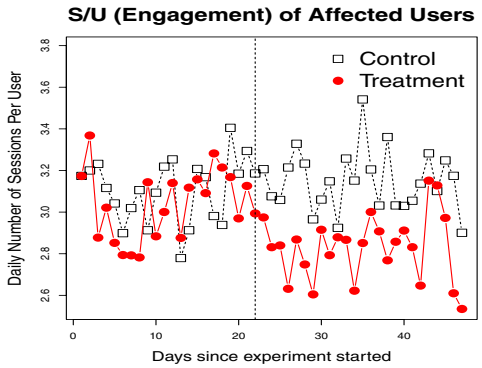


Figure 4: S/U change of the affected users in treatment and control. Users were affected on and after the 21st day of the experiment, indicated by the vertical dash line.

other non-affected users in the first three weeks, while exhibit different usage patterns after they were affected by bad relevance queries.

In our study, we empirically tried a number of k's for the affected query set but due to space limitation only $k = 0.6$ is reported here. We ended up having roughly 450 queries in the set. After filtering, 5,134 and 5,287 users were selected from treatment and control groups, respectively.

Figure 4 illustrates the change of user engagement in terms of S/U over time, where the vertical grey dotted line indicates the affected date[2]. During the first three weeks, no significant difference can be observed for the treated users. Nevertheless, once treated users were affected, we immediately notice substantial drop of engagement rate. The value of $\%\Delta(S/U)$ suggests a 5% drop during the first few affected days, and raises up to as high as 20% at peak.

On the other hand, UQ/U shows a similar pattern as S/U for the first three weeks, as demonstrated in Figure 5. However, as soon as treated users were affected, the number of

[2]**Due to the sensitivity of data, all the absolute values reported in this paper have been linearly scaled.**

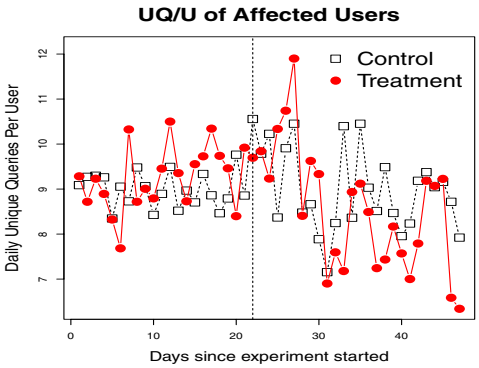


Figure 5: UQ/U change of the affected users in treatment and control. Users were affected on and after the 21st day of the experiment.

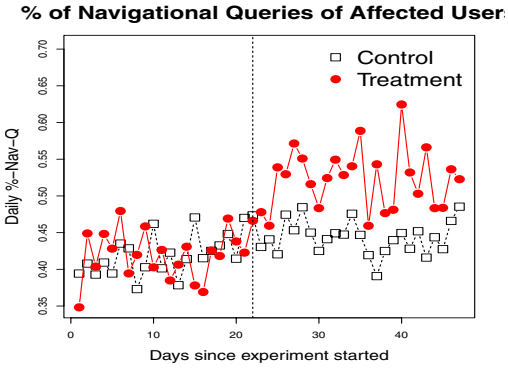


Figure 6: Query distribution change over time. Y-axis indicates the percentage of navigational queries users issued.

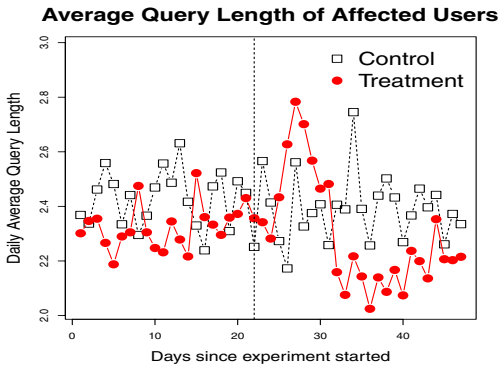


Figure 7: Average daily query length for affected users. Shorter queries are more likely navigational while longer queries are tail/hard queries.

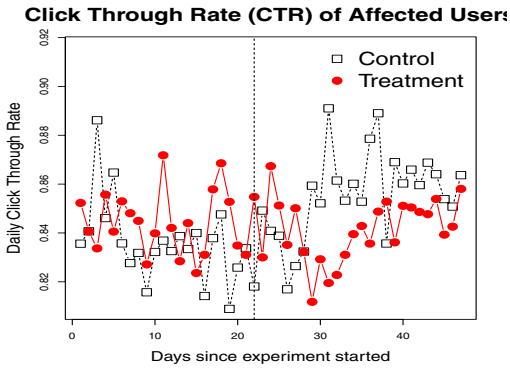


Figure 8: Change of CTR of impacted users. CTR first went up after the 21 day of experiment then suddenly dropped significantly.

unique queries they issued clearly went up for the next 6-7 days, which gradually decreased in the following weeks as UQ/U in treatment became less than that of controlled users. This further confirms the same findings as we found in the previous section: after initially affected by a relevance engine, users spent more effort on refining their queries to complete the same task, but gradually lost confidence so they came back less frequently.

Now that we know that user engagement decreases when exposed to a bad relevance engine, we want to further understand how that influence users' queries and clicks over time. To start with, we classify user queries into either navigational or informational [2, 12] as discussed in previous section. The distribution of query types is depicted in Figure 6. Before affected, both control and treatment users have roughly the same % of navigational queries — around 40% daily. After the affected point, we can clearly observe a soar of navigational queries merely after three days for those who got affected, which raised much as 64% on day 40.

We then examine the change of user query length. Previous research has shown that there is a high correlation between query length and its difficulty [17], i.e., navigational queries are mostly short queries (1 or 2 words), while tail queries are longer than head queries in general. With the relevance getting worse in treatment, we expect users to issue easier queries than before. This assumption is confirmed in Figure 7. After initially affected, users started to reformulate their queries and therefore a sudden increase of query length is observed for the first week after affected. Users then gradually became inpatient and only issued short queries.

Last but certainly not the least, we quantify the changes in user clicks in terms of click through rate (CTR). In general, higher CTRs correspond to better user satisfaction. From Figure 8, we notice that CTR dropped substantially during the first week after treated users were affected, as they began to issue more queries to fight against bad relevance but eventually failed. Later, as those users started to issue more navigational queries, the CTRs gradually increased to a comparable rate to those of the controlled users.

We summarize these five findings for affected users in Table 2. Statistical significance test is conducted where the changes of the five metrics are all significant. Figure 10 also plots the $\%\Delta$ changes for these five metrics on daily basis. Note that before the affected date, most of the date points are within $[-0.1, 0.1]$ range, meaning there is no big differences between treated and controlled users. On the other hand, after affected, we observe a lot of escalated points: treated users have as much as 40% more navigational

Metric	Before Affect	After Affect	$\%\Delta(f)$
S/U	3.0347	2.8238	7.47%*
UQ/U	9.2907	8.4752	9.62%**
% Nav Q	0.4237	0.5247	-19.25%**
Avg Q Len	2.3270	2.2908	1.58%*
CTR	0.8438	0.8356	0.98%*

Table 2: Summary of changes for the 5 metrics discussed before and after affected. Numbers here are for the affected users in the treatment group. *: p-val <0.05. **: p-val <0.01. *Values are linearly scaled due to sensitivity.*

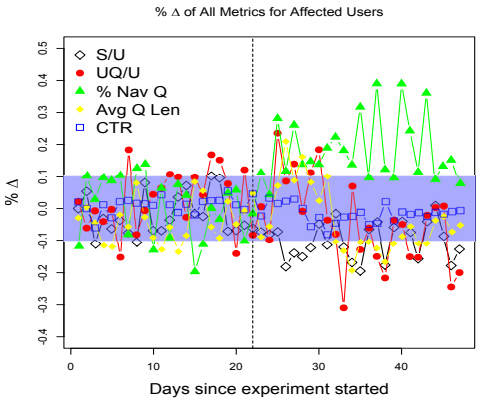


Figure 10: Daily changes for the 5 metrics before and after the affected date. Most metrics are not significant before affected date but turned out to be significant afterwards.

queries, 20% and 30% less daily sessions (S/U) and unique queries (UQ/U), respectively.

Examples of sessions are shown in Table 1 between treated and controlled users, where users started with the same query and ended up clicking on the same result. With these constraints, we assume that users have the same search intent. However, in these examples, we see that users in treatment spent significantly more effort in completing the same tasks, by issuing and reformulating more queries and clicking on more results. Figure 9 illustrates a randomly-chosen individual user whose engagement dropped during the experiment, where the y-axis shows the number of daily sessions. It is obvious that before affected, the user tended to issue more complicated informational queries. After affected, we observe that most of the queries were navigation-only queries. The S/U for that user dropped from 4.3 to 2.2, significantly.

4.3 Degree of Impact Analysis

In this section, we want to further confirm our findings in the previous section by dividing treated users into two buckets: heavily-affected users and lightly-affected users. Our assumption is that users who are exposed more often to bad relevance results should demonstrate stronger signals to the metrics we discussed, than those gently affected.

Figure 11 shows the cumulative distribution function (CDF) for users by affected queries. Since over half of the users

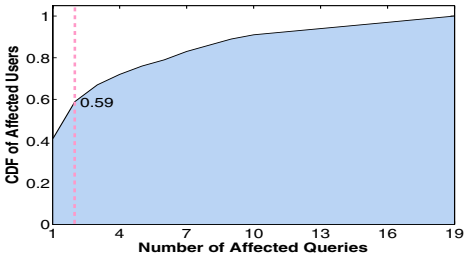


Figure 11: Cumulative Distribution Function of users in terms of affected queries.

Metric	Lightly-Affected	Heavily-Affected
S/U	2.8408	2.8077
*UQ/U	8.4871	8.4421
**% Nav Q	0.4929	0.5713
*Avg Q Len	2.3007	2.2863
*CTR	0.8407	0.8258

Table 3: Breakdown of metrics in two buckets of users based on the degree of affect. Except for S/U, all other metrics show statistical significance.

were only affected by two or less queries, we use two as our cut-off to separate users into heavily-affected and lightly-affected users, respectively.

The comparative results are shown in Table 3. In general, our assumption is confirmed by the data: heavily-affected users exhibited stronger signals than lightly-affected users, with five metrics all pointing to the correct direction. Except for the S/U metric, the rest are all statistically significant. Specifically, heavily-affected users issued substantially more navigational queries (57.13%) than the other group (49.29%). The CTR of heavily-affected users also dropped from 0.84 to 0.82. Also, heavily-affected users are less likely to issue new queries as indicated by UQ/U (8.44 vs. 8.48).

5. PREDICT USER ENGAGEMENT CHANGE

For search engines, keep the users engaged is the key to their success. Now that we understand how users behave under the circumstance of bad relevance, we want to further leverage these signals to predict the change of user engagement in the future. To start with, we formulate this problem as a binary classification task using machine learning technique. Our primary objective is to quickly detect user session *decrease* in practical large-scale online experiments so that search engines can take actions properly.

5.1 Data Preparation

Specifically, we focus on predicting user *weekly* number of sessions, whether decrease (-) or increase (+). Due to the fact of session differences during weekdays and weekends (i.e., users issue more queries during weekdays than weekends), we decide to aggregate data to weekly level so that this weekday/weekend impact is minimized. However, we believe that our work can be easily adapted to predict daily session changes with the addition of some daily-based user features.

There are various ways to formulate the prediction prob-

Treatment		Control	
query	**click**	**query**	**click**
doc bao bao daily express doc bao express	docbao.com.vn www.express.co.uk **vnexpress.net**	doc bao	**vnexpress.net**
free credit report free annual transunion credit report federal free annual credit report	www.annualcreditreport.com /cra/index.jsp **www.ftc.gov/bcp/edu/** **microsites/freereports/**	free credit report	**www.ftc.gov/bcp/edu/** **microsites/freereports/**
grammy awards 2011 grammy awards 2011 live performances at grammys 2011 grammys 2011 performers	 idolator.com/5766501/grammy- awards-2011-performances **www.grammy.com/news/** **grammy-performers**	grammy awards 2011	**www.grammy.com/news/** **grammy-performers**

Table 1: Examples of sessions in comparison. Users in treatment spent more effort for the same query intent than controlled users. Bolded URLs are desired pages.

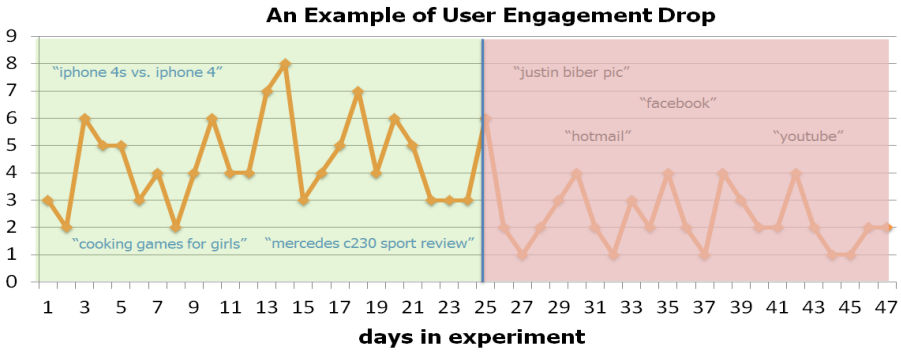


Figure 9: An example of user engagement drop before and after the user was affected in terms of daily session number. It is also clear that the user starts to issue more navigational queries after exposed to a bad relevance engine.

lem. Ideally, we would like to be aware of the engagement change as soon as possible, i.e., during the first week of the experiment. However, from a machine learning perspective of view, this approach is infeasible due to the lack of *training data*. Therefore, to make reasonable predictions, we are required to have at least one week training data to start with so that the best we can do is to predict the second week's engagement given the first week's signals. In what follows, we experiment with different amounts of training data: one with three week's training data and one with one week's training data. Since users who have three week's engagement are usually heavy users while users with one week data are often low-frequency users, these two experiments essentially measure the effect of degraded search relevance to heavy and low-frequency users, respectively.

5.2 Prediction for High-Frequency Users

From the experimental data, we filter users who have at least one activity during week 1, 2, 3 and 4 of the experiment. Data is aggregated to user level as the prediction is per use base. Prediction is made on week 4 by leveraging user data from week 1-3. A positive label is assigned if week 4's total session number is *more than* that of week 3. Otherwise a negative label is assigned. Overall, 77,940 positive instances and 54,470 negative instances are collected.

5.2.1 Classifiers

We experimented with a variety of classification algorithms, including logistic regression, boosted decision trees and linear Support Vector Machines (SVM) [4]. Among them, linear SVM demonstrated the best performance in terms of both classification accuracy and scalability. Therefore, we only report the results from linear SVM in our experiments.

5.2.2 Features

Table 4 summarizes the set of basic features we used, which covers both aspects of engagement and relevance. Some of them are closely related to the features used in the previous section. For example, *numSessions* is the same as S/U on a weekly basis; *numUnqQueries* equals weekly UQ/U; *numNoClickQueries* is correlated with CTR and etc.

We then derive a set of Δ-features based on the basic features. To be concrete, for each basic features, we calculate the Δ differences between each two weeks of week 2 to week

4. For example, for *numClicks*, it has three Δ-features for each pair of weeks: $\Delta W_3W_2numClicks$, $\Delta W_3W_1numClicks$ and $\Delta W_2W_1numClicks$, where

$$\Delta W_iW_jnumClicks = \frac{W_inumClicks - W_jnumClicks}{W_inumClicks}. \quad (2)$$

We further transform some of the count features into percentile-based features. Features like *maxSessionLength* and *QueryTimeInterval* are un-bounded so that directly apply them to classifiers may not be an optimal choice. Consequently, these features are transformed into percentile and included along with the original count features. As a result, a total of 60 features are used for training the classifier.

5.2.3 Results

Since we have over 50% more training instances in the positive class than the negative class, we perform subsampling from both classes with certain ratio to address the class imbalance issue. To be concrete, we randomly sample 20,000 instances from each class, resulting a total of 40,000 training examples, and then randomly split them on a 50/50 ratio to fit a linear SVM model. This process is repeated 10 times and the average result is reported here. To find the best parameters for the model, we perform a grid search. Overall, the best performance is reached with $s = 2$, i.e., L2-regularized L2-loss support vector classification (primal), and $c = 2$ the cost factor.

We compare to several baselines:

Baseline (Session): baseline that leverages only weekly number of sessions as the predictor.

Baseline (CTR): baseline that uses only weekly click through rate (CTR). $CTR = 1 - \frac{numNoClickQueries}{numQueries}$.

Baseline (Basic): baseline that uses the basic features only, as listed in Table 4.

Figure 12 summarizes the performance of all algorithms in comparison. It can be observed that with only the session feature, the result is almost equal to random guess around 50%, which demonstrates a very poor correlation between users weekly sessions. With the introduction of click features (CTR), the result is substantially improved. This indicates a strong connection between search relevance and user engagement. The performance is further improved by leveraging all basic features, except for a few cases at very low recalls. Finally, by combining basic features with Delta features as well as percentile features, the algorithm achieves the best predictive performance among all. Overall, linear SVM achieves 72.17% of accuracy when using all 60 features. Noticeably, at low levels of recall, by adding the Δ-based features and percentile-based features, the model is capable of improving the precision by 10% to 12% over basic features.

We list the top-15 highest weighted features in Table 5 from the linear SVM model. Overall, *numClicks*, average *avgSessionLength*, and *numNoClickQueries*, as well as their Δ-based features, play the most important roles in the model. We also notice that *numSessions*-related features did not make into this list, which again demonstrates a poor correlation between users weekly session numbers.

5.3 Prediction for Low-frequency Users

Given the fact that user's last weekly features (W_3 in our case) are the most important features as shown in Table 5, here we build another model by leveraging only one week features. A legitimate reason for doing this study is due to the fact that lots of search engine users are low-frequency users, whose activities cannot be tracked consistently by search engines for many reasons (e.g., a user who clears its browser cookies will have a new user id). Therefore, we would like to see how our algorithm performs when the data is sparse.

To this end, we extract a list of users from our data who have at least one activity in each of two *consecutive* weeks. We then randomly sample 20,000 users from the list. Features are extracted from the first week to make prediction for the second week. The data is split into 50/50 for training and testing. Note that since we only have one week data, we are unable to construct Δ-based features. Therefore, only basic features and percentile features are used for training.

Figure 13 demonstrates the result of precision-recall curves for using one-week feature, as well as the model that leverages all features for comparison. Comparatively, our algorithm is able achieve 68.59% of accuracy for low-frequency users even with one-week data, which is slightly worse (∼4%) than the model using all features.

Basic Features
numSessions: total number of sessions
numQueries: total number of queries
numUnqQueries: total number of unique queries
numNavQueries: total number of Navigational queries
numNoClickQueries: total number of queries without clicks
numClicks: total number of clicked URLs
avgQueryLength: average length of user queries
avgClickPosition: average SERP position of clicks
avgSessionLength: average length of sessions (# of queries)
maxSessionLength: maximum length of sessions
QueryTimeInterval: average time gap between two queries
ClickEntropy: the entropy of the user's all clicks

Table 4: The basic features used in this paper. All features are weekly-based. These features are used to form a total of 60 features used by the classifier.

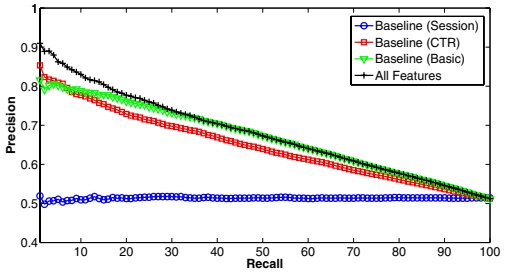


Figure 12: Precision-Recall curve for all algorithms.

6. DISCUSSION AND CONCLUSION

From our longitudinal analysis, we have observed that the change of user engagement follows a complex trend that is affected by multiple factors, therefore sometimes counter-intuitive. One would assume that under a deliberate setback of search relevance, user's engagement should immediately drop, which is, however, different from what we have observed. Our analysis indicated that user engagement in a short-term actually increased significantly, due to the fact that users tend to spend more effort by issuing more refor-

Feature Name	Weight
$\Delta W_3 W_2 numClicks$	0.041983
$W_3\_avgSessionLength$	0.034985
$\Delta W_3 W_1 numQueries$	0.025762
$W_3\_numClicks$	0.020579
$\Delta W_3 W_2 numNavQueries$	0.019042
$\Delta W_2 W_1 numQueries$	0.019013
$W_2\_avgQueryLength$	0.015908
$W_2\_numClicks$	0.015897
$\Delta W_3 W_1 numNoClickQueries$	0.014183
$\Delta W_2 W_1 avgSessionLength$	0.014054
$W_1\_avgSessionLength$	0.011007
$W_1\_numNoClickQueries$	0.010058
$\Delta W_3 W_1 avgSessionLength$	0.008692
$W_3\_avgQueryLength$	0.008319
$\Delta W_2 W_1 numNoClickQueries$	0.003609

Table 5: Top 15 highest-weighted features.

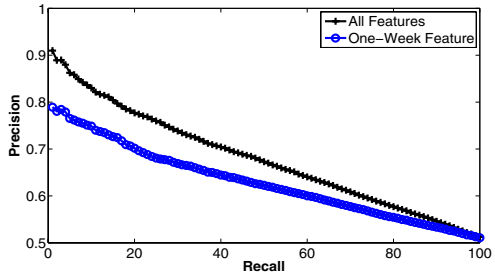


Figure 13: Precision-Recall curve for low-frequency users.

mulation, to accomplish their search tasks. As a result, during the early stage of our experiment, engagement metrics are indeed *negatively correlated* with search relevance. Nevertheless, we did observe that engagement finally dropped after a certain period of time, when users eventually gave up trusting the current search engine. Our further deep-dive analysis demonstrated that after users lost their momentum, they tend to issue more navigational-type queries (e.g., "facebook", "amazon"), and have less queries in a session, as well as a substantial drop of the click through rate.

By isolating affected users from normal users, our time series study focused on a set of core metrics that defines user behavioral change over time. With the help of these user-level and session-level features, we proposed a machine learning framework that predicts user's weekly engagement change. The model achieved over 72% of accuracy for high-frequency users and 66% for low-frequency ones. In practice, predicting user engagement is indeed a difficult task. And therefore, it is reasonable for one to assume that the number of sessions a user is going to issue in the future week is highly correlated to her current week's number, as well as the trend in the past few weeks. However, our model revealed that the highest correlated feature with engagement was in fact the number of clicks, which turns out to be a relevance metric, while the number of sessions did not make into our Top-15 feature list. This finding supports the correlation between search relevance and user engagement in a positive way.

We believe ourselves to be among the first to study the relationship between engagement and relevance under the setting of a deliberate relevance setback. Commercial Web search engine companies like Google and Bing often perform numerous online A/B experiments before they finally decide to ship new features to customers. In terms of search relevance, some new algorithms may indeed improve user satisfaction while others may eventually fail. The ship or no-ship decision is often depend on how well those online metrics perform like the ones we studied in the paper. Our study revealed that even algorithms with bad relevance may still be able to see a positive signal during the early stage of the experiment. Therefore, our advice is *not to celebrate too early* even if the signals look very positive in the first few days. We should consider keeping the experiments running for a fairly reasonable amount of time (e.g., in our case at least for two weeks), until the signals become stable or a clear trend is observed.

Our study in this paper mainly targeted the relevance domain. It would be interesting to see if the same methodology applies to other domains, e.g., User Interface (UI) changes. In the future, we also plan to further improve the accuracy of our prediction model by incorporating more features.

7. ACKNOWLEDGEMENTS

We would like to thank Pavel Dmitriev, Ya Xu, Brian Frasca, Fritz Behr, Bing Data Mining Team and Bing Relevance Team for their numerous support for this project.

8. REFERENCES

[1] E. W. Anderson and M. W. Sullivan. The antecedents and consequences of customer satisfaction for firms. *Marketing Science*, 12(2):125–143, 1993.

[2] A. Broder. A taxonomy of web search. *SIGIR FORUM*, 36(2):3–10, 2002.

[3] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *CIKM '11*, pages 611–620, 2011.

[4] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9(6/1/2008):1871–1874, 2008.

[5] H. A. Feild, J. Allan, and R. Jones. Predicting searcher frustration. In *SIGIR '10*, pages 34–41, New York, NY, USA, 2010. ACM.

[6] V. Hu, M. Stone, J. Pedersen, and R. W. White. Effects of search success on search engine re-use. In *CIKM '11*, pages 1841–1846. ACM, 2011.

[7] S. Ieong, N. Mishra, E. Sadikov, and L. Zhang. Domain bias in web search. In *WSDM '12*, pages 413–422, New York, NY, USA, 2012. ACM.

[8] B. J. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *WWW '07*, pages 1149–1150. ACM, 2007.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[10] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *KDD '07*, pages 959–967.

[11] Z. Liao, Y. Song, L.-w. He, and Y. Huang. Evaluating the effectiveness of search task trails. In *WWW '12*, pages 489–498, New York, NY, USA, 2012. ACM.

[12] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[13] M. Meiss, J. Duncan, B. Gonçalves, J. J. Ramasco, and F. Menczer. What's in a session: tracking individual

behavior on the web. In *HT '09*, pages 173–182, 2009.

[14] B. Mittal and W. M. Lassar. Why do customers switch? the dynamics of satisfaction versus loyalty. *Journal of Services Marketing*, 12:177–194, 1998.

[15] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *WWW '12*, pages 599–608, 2012.

[16] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *CIKM '08*, pages 43–52, New York, NY, USA, 2008. ACM.

[17] Y. Song and L. He. Optimal rare query suggestion with implicit user feedback. In *WWW '10*, pages 901–910, New York, NY, USA, 2010. ACM.

[18] J. E. R. Staddon. *Adaptive Behavior and Learning*. Cambridge University Press, 1983.

[19] D. Tang, A. Agarwal, D. O'Brien, and M. Meyer. Overlapping experiment infrastructure: More, better, faster experimentation. In *CIKM 2010*, pages 17–26.

[20] K. Wang, T. Walker, and Z. Zheng. Pskip: estimating relevance ranking quality from web search clickthrough data. In *KDD '09*, pages 1355–1364, 2009.

[21] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *WWW '07*, pages 21–30, New York, NY, USA, 2007. ACM.