Impact of Response Latency on User Behavior in Web Search

Ioannis Arapakis
Yahoo Labs
Barcelona, Spain
arapakis@yahoo-inc.com

Xiao Bai
Yahoo Labs
Barcelona, Spain
xbai@yahoo-inc.com

B. Barla Cambazoglu
Yahoo Labs
Barcelona, Spain
barla@yahoo-inc.com

ABSTRACT

Traditionally, the efficiency and effectiveness of search systems have both been of great interest to the information retrieval community. However, an in-depth analysis on the interplay between the response latency of web search systems and users' search experience has been missing so far. In order to fill this gap, we conduct two separate studies aiming to reveal how response latency affects the user behavior in web search. First, we conduct a controlled user study trying to understand how users perceive the response latency of a search system and how sensitive they are to increasing delays in response. This study reveals that, when artificial delays are introduced into the response, the users of a fast search system are more likely to notice these delays than the users of a slow search system. The introduced delays become noticeable by the users once they exceed a certain threshold value. Second, we perform an analysis using a large-scale query log obtained from Yahoo web search to observe the potential impact of increasing response latency on the click behavior of users. This analysis demonstrates that latency has an impact on the click behavior of users to some extent. In particular, given two content-wise identical search result pages, we show that the users are more likely to perform clicks on the result page that is served with lower latency.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Web search engine; response latency; user behavior

1. INTRODUCTION

The vast majority of research conducted so far in the information retrieval field has focused on facilitating and improving the engagement of end users with search systems. To this end, the related research tackled a number of side problems, such as query recommendation, snippet generation, and result presentation. The core research, however, has been on improving the quality of results served by such systems with the eventual goal of satisfying the information needs of users (commonly expressed as short keyword queries). In these works, the main quality metric has been the degree of relevance between the served results and the user query while other metrics (e.g., result diversity or recency) have attracted some attention as well.

Improving the quality of search results often required coming up with sophisticated or costly solutions (e.g., storing more information in the inverted index or using machine-learned ranking strategies), thus increasing query processing times. This, when coupled with the continuous growth of the indexable Web and the ever-increasing query volumes of commercial search engines in the last two decades, shifted some research attention to the efficiency of search systems. This line of research was often orthogonal to the aforementioned research on search result quality. In a large number of studies, proposed efficiency improvements resulted in an interesting trade-off between the speed of a search system and its result quality, which is used as a proxy for quantifying user satisfaction. Most often, these studies aimed to reduce the response latency of a search system with little or no sacrifice in result quality. But, the impact of speed improvements on the end user behavior and experience was not investigated in detail.

In practice, serving search results at the right speed is of vital importance to a commercial web search engine. Serving search results too slow or too fast both may result in certain financial consequences for the search engine. On the user side, the new generation of web users are impatient and have limited time. They expect subsecond response times from a search engine upon submission of their queries. High response latency is known to distract users and cause them to issue fewer queries than usual, decreasing users' engagement with the search engine in the long term [18]. This, in turn, can make a negative impact on the advertising revenue of the search engine. On the search engine side, commercial web search companies are known to make major investments in hardware infrastructures to cope with the growth of the Web as well as the growth of their user bases and query volumes, essentially trying to maintain their query response times at reasonable levels [5]. These investments incur a financial burden on search engine companies and may even result in financial losses if the reduction attained in query response times due to these investments does not have any positive impact on the search experience of users.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

In this paper, we are interested in the user side of the problem. Our main objective is to understand the potential impact of response latency on users' search behavior. In particular, our work aims to answer questions of the following kind. What are the main cost components in the response latency of a web search engine? At what point do added delays in response latency become noticeable by users? What is the effect of increasing response latency on the click behavior of users?

The contributions of our work can be summarized as follows. We describe the dominant factors in web search response latency and demonstrate the relative importance of each factor using real-life data traces. We conduct a small-scale, controlled user study, which reveals the differences in the way users perceive the latency. We conduct a large-scale analysis using a query log obtained from Yahoo web search, providing certain insights about the impact of increasing response latency on the click behavior of users.

The selected findings of our work are the following.

- Query processing and result page rendering times are the two main components in the response latency of a web search engine. Network latency becomes more pronounced as the end-to-end latency increases.
- The users of a fast search system are more likely to notice added delays than the users of a slow system.
- As long as the delay added to a response remains under 500ms, users cannot distinguish between a delayed response and a regular response with no added delay. When the introduced delay is larger than 1000ms, users are highly likely to notice the presence of delay.
- Given two content-wise identical search result pages, users are more likely to perform clicks on the result page that is served with lower latency.

The rest of the paper is organized as follows. Section 2 contains a brief summary of related work. In Section 3, we provide some initial experiments aiming to characterize the response latency of a web search engine. The details and findings of our controlled user study are presented in Section 4. In Section 5, we present our large-scale query log analysis. We conclude the paper in Section 6.

2. RELATED WORK

Cost of searching. A related line of research has investigated the trade-off between the cost of searching and the user effectiveness in interactive information retrieval. In recent work, the querying cost was typically represented by the physical or mental effort spent by the users when searching for certain information in a retrieval system [19]. In [1], the microeconomic theory was applied to interactive information retrieval, and it was shown that useful information obtained by a user during a search session is functionally related to the effort spent by issuing queries and examining retrieved results. In [2], the authors conducted a user study where participants were split into three groups to use different search interfaces, each requiring a different amount of physical and mental effort for issuing queries. Although most results reported by the study were not statistically significant, the authors observed that the participants who used the search interface with high querying cost submitted fewer queries, examined more result documents per query, and spent more time on search result pages. In [3], the authors simulated interactive search sessions assuming a desktop PC scenario, where querying effort is low, and a smart phone scenario, which requires high querying effort. They showed that the user effort spent on searching, when coupled with a time constraint on the session duration, affected the user experience in both scenarios. In particular, they found that the smart phone scenario led to deeper result scanning while the desktop PC scenario favored better queries.

Metrics. Certain effectiveness metrics, such as DCG [11] and RBP [15], incorporated the user effort implicitly by decaying the information gain with increasing rank (assuming users scan search results from top to bottom and spend a fixed amount of effort when examining each result). The time-based gain measure in [20] incorporated the user effort more explicitly by using the time spent scanning the results.

Page load time. There have been quite a few studies on response time of general computer systems in the context of human-computer interaction. The reader may refer to [6] for a discussion of those studies. In the more specific context of web systems, earlier studies investigated the impact of page load time on the information searching behavior of users [7, 10, 21]. The study in [21] (follow-up work to [7]) reported web page load time tolerable by users who are seeking information in the Web to be in the 7 to 11 seconds range. The same research showed that there is a latency threshold at which users start examining the content of web pages more thoroughly before navigating to new pages. Although the context is different, this finding is consistent with the cost-interaction hypothesis, which states that users examine search results in more depth before issuing queries when the querying effort is high [2]. Despite being outdated, [16] provides extensive references to studies on identifying the largest page load time that users can tolerate.

Query response latency. In [18], the authors exposed a commercial search engine's users to response time delays of varying magnitude and observed the impact of different levels of delay on users' long-term search behavior. They observed that the users who were exposed to higher time delays issued fewer queries than they usually do. Interestingly, the effects were shown to be persistent in the long-term even after the response latency had returned to the original levels. Our work differs from [18] in two ways. First, our user study allows us to introduce artificial response time delays on the client side, whereas [18] relies on server-side delays. This lets us work with more realistic (user-perceived) latency values and provides better control on certain parameters. Second, in our query log analysis, we focus on the short-term click behavior of individual users, instead of the change in aggregate query volumes, which is the main metric in [18]. The most relevant work to ours is the user study presented in [4], although it differs significantly with respect to the adopted methodology. In [4], the participants interacted with two simple interfaces serving search results at controlled latency values, and stated their preferences between a slow and a fast search interface through a questionnaire. The findings of the study regarding the impact of latency on users' preferences were mainly inconclusive. In our user study, instead of assigning participants into two fixed latency buckets, we expose each participant to multiple levels of latency, allowing us to investigate the way they perceive the latency better. Moreover, we experiment with much lower latency levels, which are more realistic for today's web standards (our latency values range between 0 and 2750ms with an increment of 250ms, whereas the latency values used in [4] range between 1 and 5 seconds with an increment of 1 second).

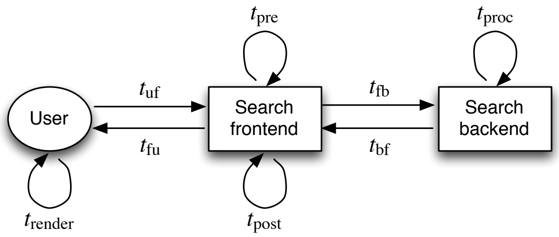
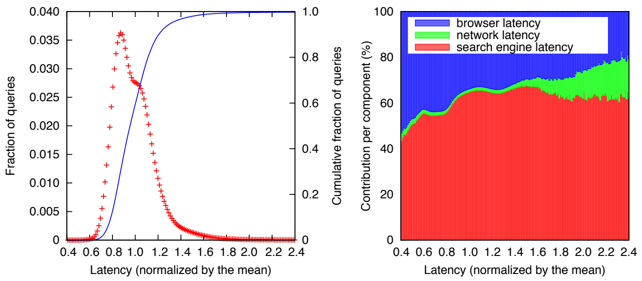


Figure 1: Latency components in web search.



(a) Distribution of latency. (b) Dissection of latency.

Figure 2: Characteristics of the response latency.

3. PRELIMINARIES

Retrieving search results. In a typical web search scenario, a user submits a query to a search engine by typing one or more keywords into a search box. The query is then transferred over the network, from the user's device to a frontend system in the web search engine. If the results of the query are already cached in the frontend system, they can be immediately served by the cache. Otherwise, the query is transformed into an internal representation after some preprocessing (e.g., query expansion, spell correction) and communicated to one or more backend query processing systems. Each backend system identifies the best-matching search results for the query by processing it on an inverted index (potentially coupling the process with machine-learned ranking). The results returned by different backends are aggregated into a final search result page, which is cached in the frontend system and communicated back to the user over the network. Finally, the search results received by the user's device are rendered using a browser. The basic search process and individual latency components are illustrated in Fig. 1.

Constituents of user-perceived response latency. In the aforementioned process, user-perceived response latency is defined as the time difference between the rendering of retrieved search results in the user's browser and the submission of the query. This end-to-end latency involves three main components: network latency, search engine latency, and browser latency. The network latency is composed of the round-trip time between the user and the web search engine frontend ($t_{\text{uf}} + t_{\text{fu}}$). This latency is known to correlate well with the physical distance between the user and the search engine and, to some extent, with the available network bandwidth. The search engine latency corresponds to the time difference between the arrival of the query to the search engine and the start of results' transfer to the user ($t_{\text{pre}} + t_{\text{fb}} + t_{\text{proc}} + t_{\text{bf}} + t_{\text{post}}$). Finally, the browser latency corresponds to the time difference between the reception and rendering of search results in the user's browser (t_{render}).

Characterizing response latency. Fig. 2(a) shows the distribution of response latency values observed in Yahoo web search (please refer to Section 5.1 for the details of the query log used).[1] According to the solid curve in the figure, about 60% of queries are answered under the mean latency and about 99% of queries have a response latency less than 1.8μ. The latency distribution (dotted curve) is observed to have a slight distortion after the peak. This behavior is because the distribution actually involves two sub-distributions with peaks around 0.85μ and 1.05μ. The former sub-distribution is due to queries served by the result cache while the latter is due to queries processed in an actual backend search system.

Fig. 2(b) shows the contribution of different latency components to the user-perceived response latency. We observe that the end-to-end latency is mainly determined by the search engine latency and the browser latency. While the contribution of the two latency components are similar when the responses are fast (e.g., around 0.5μ), the search engine latency becomes the dominant factor as the response times increase. At much larger latency values (e.g., around 1.6μ), the network latency starts to become more noticeable.

Possible experimental methodologies. There are three possible experimental methodologies one can adopt to carry out a study in our context: bucket testing, controlled user study, and query log analysis. In case of bucket testing, the users of the search engine may be split into buckets, each subject to a different set of test parameters (e.g., varying added delays as in [18]). Bucket testing enables a large-scale and real-life study. However, it is not easy to control certain parameters and observe the real user experience. User studies are typically much smaller in scale [4], but a wider range of parameters can be explored in a controlled manner. The downside is the difficulty of generalizing the findings. Finally, query log analysis may let us make observations using recorded search behavior of users. This kind of an analysis can be large scale, but has little flexibility for introducing new parameters. In our work, we adopt the user study (Section 4) and query log analysis options (Section 5). We leave bucket testing as a future work.

4. CONTROLLED USER STUDY

To demonstrate the impact of response latency on search behavior we carried out two controlled experiments that examine users' interactions with two different search sites. The first study examines users' sensitivity to different levels of latency as well as their perception of response time. The second study demonstrates the effects of increasing response latency on the search experience and, more specifically, on user engagement and satisfaction. As a side contribution, we also looked at potential bias due to search site branding.

4.1 User Sensitivity to Latency

Experimental design. The experiment used a repeated-measures design with two independent variables: search latency (with 12 levels in milliseconds: "0", "250", "500", "750", "1000", "1250", "1500", "1750", "2000", "2250", "2500",

[1] Due to the confidential nature of the data, we normalize reported response latency values by the mean latency (μ).

"2750") and search site speed (with two levels: "slow", "fast"). The search latency was controlled by using a client-side script that adjusted search latency by a desired amount of delay. The search site speed was controlled by using either a commercial search site with a generally slow response rate ($\mathrm{SE}_{\mathrm{slow}}$) or a commercial search site with a generally fast response rate ($\mathrm{SE}_{\mathrm{fast}}$). Although the two search sites were different, the returned search results were very similar due to the nature of queries used (see Procedure). The dependent variables were (i) sensitivity to search latency and (ii) prediction accuracy of search latency.

The scatter plot in Fig. 3 shows the response latency values observed for $\mathrm{SE}_{\mathrm{slow}}$ and $\mathrm{SE}_{\mathrm{fast}}$ upon submission of identical queries. We observe $\mathrm{SE}_{\mathrm{slow}}$ to be somewhat slower than $\mathrm{SE}_{\mathrm{fast}}$. For almost any query, $\mathrm{SE}_{\mathrm{fast}}$ has lower latency.

Apparatus. In our experiment, we used a desktop computer equipped with a 24″ LCD monitor, keyboard, and mouse. In the background, we ran a custom-made javascript that controlled the search latency. The script was deployed using the Greasemonkey[2] extension in a Mozilla Firefox web browser. It captured a series of browser events (e.g., mouseover, click, or keypress) and logged the unix timestamps for every query submitted and each search engine result page (SERP) rendered in response to a query.

Questionnaires. At the beginning of the study, the participants were asked to fill in an entry questionnaire, which gathered background and demographic information, as well as information about their previous experience with online search. A set of scales was developed specifically for our study (e.g., easy/difficult, relaxing/stressful, and satisfying/frustrating) based on users' response to the statement "Using a search site is generally...".

Participants. There were 12 participants (female=6, male=6) aged from 24 to 41 and free from any obvious physical or sensory impairment. The participants were of mixed ethnicity (Catalan, Chinese, Italian, German, Greek, Korean, Persian), came from a variety of educational backgrounds (41.6% had an MSc degree and 58.3% had a PhD degree), and were all proficient with the English language (8% intermediate level, 75% advanced level, 17% native speakers). They were primarily pursuing further studies while working (54.3%) although there were a number of students (33.3%) and full-time employees (16.6%). Participants reported using a search site at home or work very often ($M = 6.58, SE = .79$). In addition, they indicated that they find online searching a very easy ($M = 6.00, SE = 1.53$) and somewhat satisfying ($M = 5.50, SE = 1.16$) task.

Procedure. The user study was carried out in a laboratory setting and followed a think-aloud protocol. At the beginning of each session, the participants were informed about the conditions of the experiment and were asked to complete a demographics questionnaire. Each participant then performed two tasks. Both tasks involved submitting a fixed number of randomly selected navigational queries, i.e., queries that seek a single website or web page of a single entity (the web domain list was created using the web analytics provided by Alexa[3]). We limited the study to navigational queries because they impose a smaller cognitive load to the searcher (compared to other types of queries), promote a convergence in the search intent across all users, and do not require native-level knowledge of the English language. Therefore, by mitigating the effort of query formulation, our participants were able to assess the latency effect better.

The first task asked the participants to report to the experimenter their subjective impression of the search site's response latency after each submitted query, i.e., whether they felt that the response was "slow" or "normal". In this task, the search latency was increased by a fixed amount that ranged from 0 to 1750ms, using a step of 250ms. Each latency value (0ms, 250ms, ..., 1750ms) was introduced five times and in a random order, in combination with 40 randomly selected navigational queries. The provided navigational queries were submitted to the search site the same way they would be submitted in a realistic search scenario, i.e., through typing and clicking.

The second task required the participants to provide an estimation of the search latency in milliseconds for each submitted query. Participants were instructed to consider as search latency the time from the query submission until the SERP was rendered. The search latency was set to a fixed value that ranged from 500ms to 2750ms, using a step of 250ms. Similar to the previous task, each latency value was introduced five times and in a random order, in combination with 50 navigational queries. To familiarize themselves with the default behavior of the search site and establish a measure of comparison, the participants were asked to submit a set of training queries before each task. Finally, to control for order effects, the task assignment was randomized.

Results (first task). Fig. 4 shows the distribution of cases where the participants felt that the response was slow or normal. Based on that plot, Fig. 5 shows the likelihood that the participants will feel the added delay in response time. In case of $\mathrm{SE}_{\mathrm{fast}}$, when there was no added delay, the participants could almost perfectly felt the absence of delay (with $1 - 0.02 = 0.98$ probability). In case of $\mathrm{SE}_{\mathrm{slow}}$, however, their accuracy was considerably lower ($1 - 0.13 = 0.87$ probability), potentially due to the high variation in response time of $\mathrm{SE}_{\mathrm{slow}}$. In general, participants could distinguish slow response with much higher likelihood when they were using $\mathrm{SE}_{\mathrm{fast}}$. For example, when the added latency was 750ms, the likelihood of participants to feel the added latency was not different than random in case of $\mathrm{SE}_{\mathrm{slow}}$, but they were able to notice the added latency with much higher likelihood (around 0.82 probability) in case of $\mathrm{SE}_{\mathrm{fast}}$. For both search engines, added delays under 500ms were not easily noticeable by participants (not better than random prediction) while added delays above 1000ms could be noticed with very high likelihood. Fig. 6 displays similar data, but this time comparing male and female participants. According to the figure, female participants are observed to be better in noticing small increases in response time than male participants. But, there is no significant difference between males and females when the added delays are large.

Results (second task). In Figs. 7 and 8, we show the predicted versus actual latency values for individual participants using $\mathrm{SE}_{\mathrm{slow}}$ and $\mathrm{SE}_{\mathrm{fast}}$, respectively. The results reveal considerable differences in the way individuals perceive the latency. In case of $\mathrm{SE}_{\mathrm{slow}}$, about half of the participants consistently overestimated the latency while the other half consistently underestimated. The prediction quality of participants have higher deviation in case of $\mathrm{SE}_{\mathrm{fast}}$ than in case of $\mathrm{SE}_{\mathrm{slow}}$. Interestingly, the average of all participants' predictions are very close to the original values in both cases.

[2] http://www.greasespot.net

[3] http://www.alexa.com/topsites

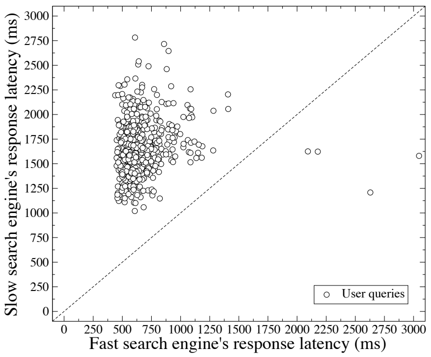


Figure 3: Response latency values attained by the fast and slow search engines for the same query.

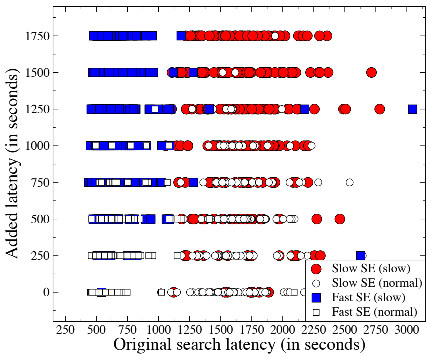


Figure 4: Participants' feelings at different levels of added latency (each point represents a query).

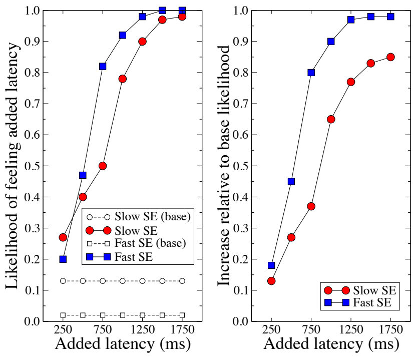


Figure 5: The likelihood of participants to feel increasing values of added latency.

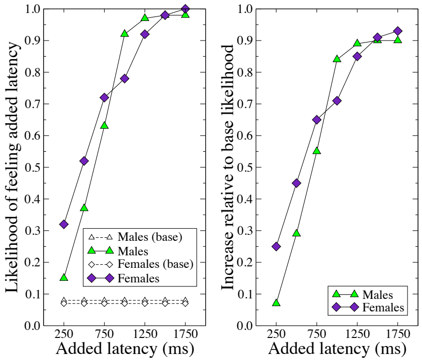


Figure 6: Impact of gender on the likelihood of feeling added latency.

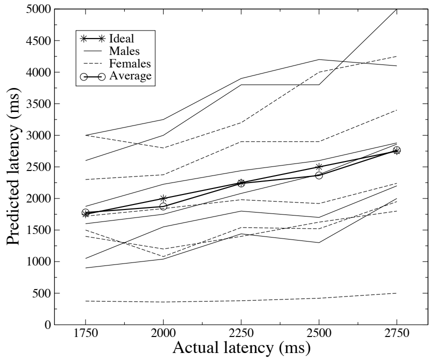


Figure 7: Latency predictions in case of the slow search engine.

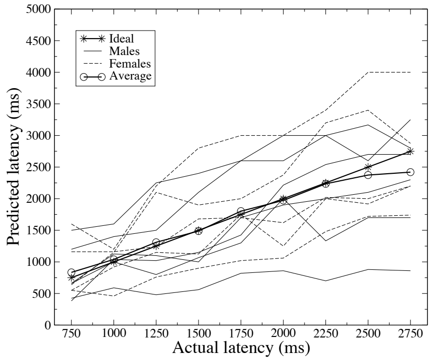


Figure 8: Latency predictions in case of the fast search engine.

4.2 Impact of Latency on Search Experience

The objective of this study is to investigate the effects of response latency on the search experience and, in particular, on user engagement and satisfaction. Two psychometric scales were used to capture hedonic and cognitive aspects of user experience: the User Engagement Scale (USE) and IBM's Computer System Usability Questionnaire (CSUQ). In addition to the psychometric scales, participants were asked to evaluate the performance and speed of the search site, as well as report the experienced frustration after each task. We speculate that, as the search latency increases, the search experience will become less engaging (i.e., low scores on all psychometric scales) and the perceived usability of the search site will be negatively impacted.

Experimental design. The experiment had a two-way, mixed design. The related measures independent variable was the search latency (with four levels in milliseconds: "0", "750", "1250", "1750"). The unrelated measures independent variables was the search site speed (with two levels: "slow", "fast"). Search latency was controlled through a client-side script that adjusted the latency by a desired amount of delay. The choice of latency values was informed by the findings from the first study (see Section 4.1). The search site speed was controlled by using either a search site with a generally slow response rate (SE_{slow}) or a search site with a generally fast response rate (SE_{fast}). Despite the two search sites coming from different brands, the returned results were almost identical due to the nature of the search queries used (see Procedure). The dependent variables were (i) experienced positive and negative affect, (ii) level of focused attention, (iii) perceived system usability, and (iv) subjective beliefs about search site performance.

Apparatus. The study used the setup in Section 4.1.

Questionnaires. We used two types of questionnaires. The first questionnaire (entry) was introduced at the beginning of the study and gathered background and demographic information, as well as information about previous experience with online search. The second questionnaire (main) was administered at post-task and included the USE and CSUQ scales. The questions were all forced-choice type and appeared in a random sequence to mitigate potential bias due to the ordering effect. The UES is multi-dimensional; its items pertain to positive and negative affect, perceived usability of the system, as well as users' felt involvement and focused attention during the task. Affect refers to the emotion mechanisms that influence our everyday interactions and can act as the primary motivation for sustaining our engagement [17] during information processing tasks or computer-mediated activities. Focused attention refers to the feeling of energised focus and total involvement, often accompanied by loss of awareness of the outside world and distortions in the subjective perception of time. The CSUQ [12] is a multi-dimensional user satisfaction questionnaire. Out of the four items it consists, we considered only the scores from the responses to system usefulness (SYSUSE). Taken

Table 1: I-PANAS-SF [22]

Positive Affect items	Negative Affect items
active	afraid
alert	ashamed
attentive	hostile
determined	nervous
inspired	upset

together, the UES and CSUQ probe users' perceptions of the pragmatic and hedonic qualities of their search interactions, as well as their perceptions of the search engine and of themselves using a technology, all of which are considered key facets of the user experience [9]. More in specific, the questionnaires inquired about the following aspects:

I-PANAS-SF. The international Positive and Negative Affect Schedule (PANAS) Short Form [22] was used to measure the affect before and after each task (Table 1). I-PANAS-SF is a validated test for measuring affect changes. It includes ten items measuring positive (PAS) and negative (NAS) affect. Participants were asked to respond on a 7-point Likert scale (very slightly or not at all; a little; moderately; quite a bit; extremely) their agreement to the statement: "You feel this way right now, that is, at the present moment", for each item. Although I-PANAS-SF may not be as efficient and accurate for capturing temporal micro-resolutions of emotional responses, there are several examples of studies from the domain of Library & Information Science [8, 13, 17] where PANAS has been successfully applied for measuring searchers' affect between search tasks. Considering that the duration of our search tasks is comparable to those in the aforementioned studies, we believe that our experimental approach to measuring emotion was reasonably accurate.

Focused attention. A 9-item focused attention subscale, part of a larger scale for measuring user engagement [17], was adapted to the context of the search tasks. The focused attention subscale has been used in past work [14] to evaluate users' perceptions of time passing and their degree of awareness about what took place outside of their interaction with the given task. Given the context of our work, focused attention was a more meaningful dimension, at least compared to other subscales of engagement (e.g., aesthetics, novelty) that were not relevant enough or were addressed by the other questionnaires employed in our study (USE, CSUQ, i-PANAS-SF). To measure focused attention, the participants were instructed to report on a 7-point Likert scale (strongly agree; disagree; neither agree nor disagree; agree; strongly agree) their agreement to each item shown in Table 2.

System usability. The CSUQ [12] was developed by IBM for measuring the perceived usability of systems in the context of realistic scenarios. A 7-point Likert scale of agreement (strongly agree; strongly disagree) that quantifies system usefulness is used for each of the 8 statements in the SYSUSE subscale. Two examples statements are "I am able to complete my work quickly using this search site" and "I am able to efficiently complete my work using this search site".

Custom statements. In addition to the USE and CSUQ-SYSUSE scales, we gathered information about the search sites' performance. We used a 7-point Likert scale of agreement for the following positive statements: (i) "This search site was fast in responding to my queries", (ii) "This search site helped me to accomplish my task in a reasonable amount of time", and (iii) "I feel satisfied with the retrieved results". Moreover, we asked our participants to indicate on a 7-point Likert scale how frustrating each search task was.

Table 2: Focused attention scale [17]

1. I forgot about my immediate surroundings while performing this search task.
2. I was so involved in my search task that I ignored everything around me.
3. I lost myself in this search experience.
4. I was so involved in my search task that I lost track of time.
5. I blocked out things around me when I was completing the search task.
6. When I was performing this search task, I lost track of this world around me.
7. The time I spent performing the search task just slipped away.
8. I was absorbed in my search task.
9. During this search task experience I let myself go.

Demographics. This study gathered the same demographics as those discussed in Section 4.1.

Participants. There were 20 participants (female=10, male=10) aged from 18 to 41 and free from any obvious physical or sensory impairment. The participants were of mixed ethnicity (Dutch, English, Farsi, French, German, Greek, Italian, Korean, Persian, Spanish, Turkish, Urdu), came from a variety of educational backgrounds (10% had a BSc degree, 50% had an MSc degree and 40% had a PhD degree), and were all proficient with the English language (10% intermediate level, 70% advanced level, 20% native speakers). They were primarily pursuing further studies while working (40%) although there were a number of students (35%) and full-time employees (25%). Participants reported using a search site at home or work very often ($M = 6.85, SE = 0.36$). In addition, they indicated that they find online searching an easy ($M = 5.75, SE = 1.91$) and somewhat satisfying ($M = 5.30, SE = 0.86$) task.

Procedure. The user study was carried out in a laboratory setting. At the beginning of each session, the participants were informed about the conditions of the experiment and were asked to complete a demographics questionnaire. Each participant had to perform four search tasks (one for each latency value). The tasks were presented in the context of a short cover story, which asked the participants to evaluate the performance of four different backend search systems. All tasks involved submitting out of a list of 200 web domains as many navigational queries as possible, within ten minutes. Participants were presented with two web browser windows: the first window displayed the search site while the second window displayed the the questionnaire. For each navigational query, participants were instructed to locate the associated URL among the first ten results of the SERP and copy-paste it in the corresponding box of the questionnaire. A set of training queries was used at pre-task to allow participants to familiarize themselves with the "default" behavior of the search site and the search task. To provide further motivation and engage the participants with the task, they were informed that a prize would be awarded to the person who will submit the most URLs in total. To control the order effects, the task assignment was randomized. Finally, the participants were randomly allocated to two search site groups, ensuring an even number of female and male participants per group.

Results. We present the findings based on 80 search tasks, carried out by 20 participants. For our analysis we

Table 3: Descriptive statistics (Mean, SD) for reported UE and CSUQ-SYSUSE scales

	SE_{slow} latency				SE_{fast} latency			
	0ms	750ms	1250ms	1750ms	0ms	750ms	1250ms	1750ms
postPAS	16.20 ± 9.04	14.50 ± 7.59	15.50 ± 7.21	15.20 ± 7.47	20.50 ± 7.82	19.00 ± 9.01	20.80 ± 9.48	19.30 ± 8.23
postNAS	7.00 ± 3.80	6.80 ± 2.70	7.60 ± 3.27	6.90 ± 3.28	6.80 ± 2.44	7.40 ± 3.03	7.40 ± 2.72	7.20 ± 2.49
postPAS-prePAS	-3.10 ± 8.49	-4.80 ± 6.46	-3.80 ± 6.34	-4.10 ± 7.11	2.50 ± 5.95	1.00 ± 6.13	2.80 ± 6.01	1.30 ± 6.29
postNAS-preNAS	0.30 ± 2.31	0.10 ± 1.10	0.90 ± 1.79	0.20 ± 2.30	-0.40 ± 2.46	0.20 ± 2.53	0.20 ± 2.74	0.00 ± 1.33
Frustration	3.20 ± 2.20	3.10 ± 2.02	2.90 ± 2.02	3.30 ± 2.21	2.80 ± 1.40	3.00 ± 1.63	3.50 ± 1.08	2.60 ± 0.84
FA	22.80 ± 9.37	22.90 ± 8.29	19.90 ± 9.26	22.20 ± 10.38	27.90 ± 13.20	26.60 ± 10.41	23.90 ± 9.23	29.50 ± 9.85
SYSUS	32.80 ± 6.73	28.90 ± 5.40	29.80 ± 7.63	27.90 ± 6.89	35.20 ± 5.35	31.30 ± 8.25	29.80 ± 8.34	33.20 ± 8.22

used several related and unrelated measures tests, like the Mann-Whitney and Wilcoxon Signed-Rank test for pair-wise comparisons, and Friedman's ANOVA for three or more conditions. Participants response to the 5-item PAS, 5-item NAS, 9-item focused attention, and 8-item CSUQ-SYSUSE scales were summed to obtain the final scores. Results are reported at a statistical significance level of .05. To take an appropriate control of Type I errors in multiple pair-wise comparisons we applied the Bonferroni correction.

Experienced affect. Table 3 (top) shows the mean scores for the positive (postPAS) and negative (postNAS) affect scale at post-task, as well as the difference Δs between the scores reported at pre- and post-task for SE_{slow} and SE_{fast}. The results indicate a decrease in positive affect for both search sites as we introduce larger latency values. The inverse effect is observed for negative affect, which increases as higher latency values are used, but this effect is more consistent in the case of SE_{fast}. None of the differences identified above were statistically significant. However, when comparing the reported postPAS and postNAS scores between SE_{slow} ($Mdn = 16.50$) and SE_{fast} ($Mdn = 21.00$) and across all latency values, the Mann-Whitney test indicated a statistically significant difference for postPAS, $U = 550.50, p < .05, r = -.31$. This small to medium effect observed for PAS between the two search sites suggests a positive bias towards SE_{fast}, despite participants having experienced the same range of added latencies. Table 3 also displays the mean scores for reported level of frustration. There were no differences among the latency values, nor between the two search sites.

Focused attention. Table 3 (middle) displays the mean scores for focused attention (FA). For the participants of SE_{slow}, the variation of the scores across the latency values does not indicate any visible trend. For the participants of SE_{fast}, we observe a decrease in small- and medium-size latencies. However, there were no significant differences between the latency values. When comparing the reported focused attention between the participants of SE_{slow} ($Mdn = 21.00$) and SE_{fast} ($Mdn = 26.00$), and across all latency values, the Mann-Whitney test indicated a statistically significant difference, $U = 568.50, p < .05, r = -.27$. This represents a small to medium effect for the focused attention observed between the two search sites. Moreover, it suggests that the participants of SE_{fast} felt more deeply involved with the search task, despite having experienced the same range of added latencies.

System usability. Table 3 (bottom) displays the mean CSUQ-SYSUSE scores per latency value and per search site. For both search sites we observe a noticeable increase in the reported usability scores. More in specific, for SE_{slow}, there was a statistically significant difference in the perceived usability of the search site depending on which amount of added latency was introduced, $\chi^2(3) = 11.00, p < .05$. Post-hoc analysis with Wilcoxon Signed-Rank test indicated a statistically significant difference in the perceived usability, as reported scores were significantly higher for latency value of "0" ($Mdn = 31.00$) compared to "1750" ($Mdn = 28.50$), $Z = -2.66, p < .008, r = -0.42$. This represents a large effect in the levels of perceived usability when search latency was increased by 1750ms. No significant differences were observed for SE_{fast}, suggesting that the participants were more tolerant towards the delays experienced for that search site despite the large latency values introduced to their search interactions. Additionally, the reported scores for perceived usability differed significantly between the participants of SE_{slow} ($Mdn = 30.00$) and SE_{fast} ($Mdn = 35.00$), $U = 596.00, p < .05, r = -.22$. Finally, none of the differences identified in the number of submitted queries per latency value were significant.

Search experience. We evaluated the search experience promoted by the two search sites by asking our participants to report their agreement to a set of custom statements. With respect to statement (i), the Friedman's ANOVA test indicated for SE_{slow} a significant difference in the perceived search site speed, depending on which latency value was added. Wilcoxon tests were used to follow up this finding but no significant differences were observed for any of the pair-wise comparisons. Furthermore, the reported perceived search site speed by participants of SE_{slow} did not differ significantly from that of participants of SE_{fast}, which is an interesting finding considering the notable difference in the search sites' performance. In regards to statement (ii), participants' belief that the search site helped them accomplish their task more quickly changed significantly over the latency values ($\chi^3 = 10.80, p < .05$). This effect was observed only for SE_{slow}. Post hoc tests revealed a statistically significant difference between the latency values "0" and "1750", $Z = -2.63, p < .008, r = -.83$. Finally, for statement (iii), none of the differences identified in the reported scores were statistically significant across the latencies and search sites.

These results help us understand that the subjective search experience may be influenced by branding, as well as users' preconceptions about the search site performance. For example, a search site perceived as "fast" or "efficient" may still result in engaging search interactions despite occasional poor performance. This suggests that a successful marketing approach could go a long way to improve the reputation of a product and positively bias the end-users.

Correlation analysis of all factors. Finally, we report the results of a correlation analysis performed across all search experience factors discussed above, including participants' prior beliefs of the search site performance. The importance

Table 4: Summary of correlations of subjective beliefs on search site performance and reported UE and CSUQ-SYSUSE scales

Beliefs	postPAS	postNAS	Focused attention	CSUQ-SYSUS	custom-1	custom-2	custom-3
SE_{slow} will respond fast to my queries	$.455^{**}$	$.041$	$.702^{**}$	$.267$	$.177$	$.177$	$.082$
SE_{slow} will provide relevant results	$.262$	$-.083$	$.720^{**}$	$.411^{**}$	$.278$	$.263$	$.232$
SE_{fast} will respond fast to my queries	$-.051^{**}$	$.245$	$.341^{*}$	$.591^{**}$	$.330^{*}$	$.443^{**}$	$.624^{**}$
SE_{fast} will provide relevant results	$-.272$	$.133$	$-.133$	$.378^{*}$	$.212$	$.259$	$.390^{*}$

*. Correlation is significant at the .05 level (2-tailed). **. Correlation is significant at the .01 level (2-tailed).

of this analysis is to understand better the influence of subjective beliefs on the hedonic and cognitive aspects on the search experience. Table 4 shows all interactions between UE and SYSUS factors, and subjective beliefs. We observe that in the case of SE_{slow}, positive bias in regards to the search site speed results in higher positive affect and focused attention, whereas strong belief that the search site will provide relevant results is positively correlated with perceived usability. On the other hand, for SE_{fast}, we observe that participants' positive expectations regarding to the search site speed is negatively correlated with positive affect and positively correlated with focused attention and perceived usability. Moreover, this favorable bias is also positively correlated with expectations that the given search site will respond fast to the queries, will be helpful in accomplishing the task in a reasonable amount of time, and will provide satisfactory results. Despite our relatively small sample, these findings suggest that search engine bias cannot be ruled out and users tend to interpret ambiguous evidence as supporting their existing beliefs. Hence, these tendencies to overestimate, or underestimate, system performance biases their interpretations of search interactions and invokes negative behaviors that may result in search site abandonment.

5. LARGE-SCALE QUERY LOG ANALYSIS

In this section, we investigate the impact of increasing response latency on the click behavior of real web search engine users. To this end, we use a random sample of web search queries obtained from Yahoo. We observe the variation in click behavior using the entire query sample as well as certain subsets of it.

5.1 Query Log and Metric

Query log. In our log, each search query is associated with various latency values measured at different steps of the retrieval process. In all of our experiments, we use the end-to-end (user-perceived) latency values. We limit our analysis to queries that originated from desktop computers in order to reduce the potential bias due to the differences in end user devices. Also, we limit the user space to the US, trying to reduce the variation in the network latency due to the geolocation of users. We select only queries issued to a particular search data center. The resulting sample after these filtering steps contains 30 million queries.

Metric. To quantify the engagement of users with retrieved search results, we use the clicked page ratio metric. Given a set of search result pages, the clicked page ratio metric is defined as the fraction of result pages where at least one result link is clicked by the user. In this metric, higher values imply that the users engage more often with the presented search results. Herein, we report results on how increasing response latency values affect this metric. Due to its confidential nature, when we display this metric in the plots, we always normalize the values by the maximum value observed in the plot. This should not form a major concern since we are more interested in the variation of the metric rather than the absolute metric values.

5.2 Impact of Latency on Click Behavior

Impact at first glance. We first try to observe how increasing response latency affects the clicked page ratio metric. To increase the granularity of measurements, we group queries into buckets at every 10 millisecond latency interval and compute the clicked page ratio metric separately for each bucket using all queries inside the bucket.

Fig. 9 shows the variation of the metric as the response latency increases. Surprisingly, instead of observing a decreasing trend, we observe two distributions with different peaks. This result is in line with the observation in Fig. 2(a). The first distribution in Fig. 9 corresponds to queries which are served by the result cache with low response times. Most of these cached queries are navigational queries, whose results are likely to receive at least one click. The second distribution corresponds to queries that are served by the relatively slow backend search system. These are mostly tail queries, which are less likely to result in a click on the results.

Intuitively, the quality of results has a considerably more important effect on the clicked page ratio metric than the response latency. In general, users are less likely to click on irrelevant results even if they are served with low response latency. On the other hand, if the results are expected to be very relevant, users may be willing to engage with the results, tolerating the high response latency.

One way to reduce the influence of result quality is to group queries according to the likelihood of their results being clicked. The underlying intuition here is that, if a query is very likely to result in a click on the search results (e.g., query "facebook"), this implies that the results are very often satisfactory for the users. In this case, any variation in the clicked page ratio is more likely to be due to changes in the user-perceived response latency. Similarly, if the results of a query rarely receive clicks, the variation in the clicked page ratio is more likely to be due to the change in the latency.

Based on this intuition, we group queries under five buckets such that the clicked page ratio of all queries within a bucket fall into the same interval. According to Fig. 10, for every query group, the clicked page ratio tends to decrease when the response latency increases, making a bottom around 1.45μ. This is probably because when the response latency exceeds a tolerable value, the users simply give up the current query, submit another query, or simply switch to another task other than searching.

Isolating result quality completely. We continue to analyze the impact of increasing response latency on the click behavior, this time trying to eliminate the influence of result quality completely. To this end, we generate all possible pairs of queries such that the query string and retrieved

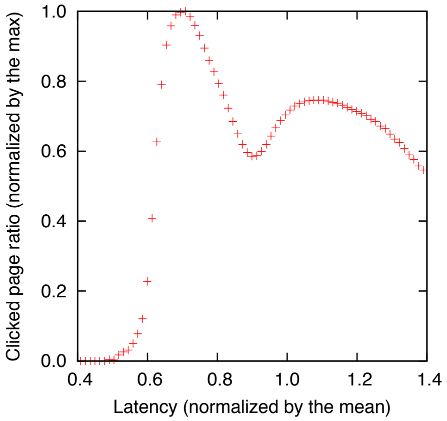


Figure 9: The clicked page ratio metric as latency increases.

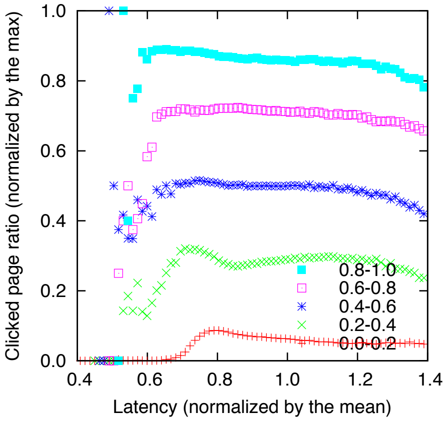


Figure 10: Clicked page ratio for five query groups created according to the same metric.

search result pages are identical for the queries in a pair. We then check whether the users preferred the result set of a query in the pair over the result set of the other query. In what follows, we refer to the query whose results were served with higher response time as the slow query while the other query in the pair is referred to as the fast query. In this analysis, we are interested in observing the following cases:

- *Click-on-fast*. At least one search result of the fast query was clicked while no result of the slow query was clicked.
- *Click-on-slow*. At least one search result of the slow query was clicked while no result of the fast query was clicked.
- *Click-more-on-fast*. At least one search result is clicked for both queries, but more results are clicked in case of the fast query.
- *Click-more-on-slow*. At least one search result is clicked for both queries, but more results are clicked in case of the slow query.

We ignore pairs of queries whose results receive the same number of clicks.

Fig. 11 shows the fraction of query pairs that fall into the *Click-on-fast* and *Click-on-slow* cases as well as the ratio between the two cases. Each data point in the plot is computed using query pairs where the latency difference of the queries in the pair is larger than the value shown on the x axis. We observe that when the latency difference increases, the fraction of the *Click-on-fast* cases increases while the fraction of the *Click-on-slow* class decreases. The ratio of the *Click-on-fast* cases to the *Click-on-slow* cases is always larger than one. This implies that, given two identical sets of search results (for two identical queries), users are more likely to click on a result retrieved with lower latency. This becomes more evident as the latency difference increases, confirming our observation in Fig. 10. Fig. 12 shows the results of a similar analysis using the *Click-more-on-fast* and *Click-more-on-slow* cases. In this case, the ratio between the *Click-more-on-fast* and *Click-more-on-slow* cases starts decreasing once the latency difference reaches 1250ms. This behavior can be potentially explained by the cost-interaction hypothesis mentioned before, i.e., clicking on search results becomes preferable to submitting new queries due to very high response latency.

6. CONCLUSIONS

We investigated the impact of increasing response latencies on user behavior in web search. To this end, we conducted a controlled user study and also performed a large-scale query log analysis. The user study revealed that up to a point (500ms) added response time delays are not noticeable by the users. However, after a certain point (1000ms), the users could feel the added delay with very high likelihood. Our query log analysis also revealed interesting findings about the change in user behavior as latency increases. In particular, given two content-wise identical search result pages, we show that the users are more likely to perform clicks on the result page that is served with lower latency.

A potential extension to our work is to repeat our study by grouping queries according to user demographics, context, and potentially many other factors (e.g., time). We believe that the subjective nature of perceived latency creates an opportunity for search engines. Search results can be served to each user at custom latencies depending on the estimated behavioral impact on the user. For example, if no negative impact is estimated on the user experience, search results may be served at high latencies by computing them using less resources. Serving results at right latencies may bring further financial benefits to search engines in the form of decreased hardware investments and reduced energy consumption. All of this, of course, requires devising certain forecasting mechanisms for accurate prediction of user-perceived response latency as well as the impact on user experience.

7. REFERENCES

[1] L. Azzopardi. The economics in interactive information retrieval. In *Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 15–24, 2011.

[2] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *Proc. 36th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 23–32, 2013.

[3] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: Simulating sessions in diverse searching environments. In *Proc. 35th Int'l ACM*

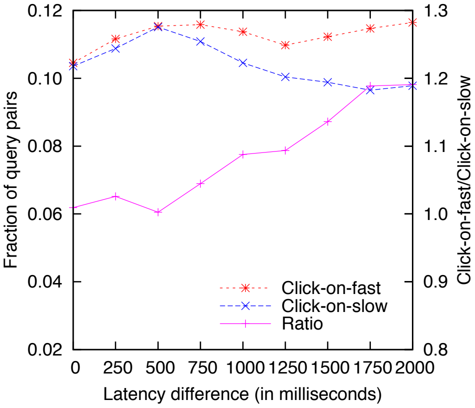


Figure 11: Removing the influence of result quality (by click presence).

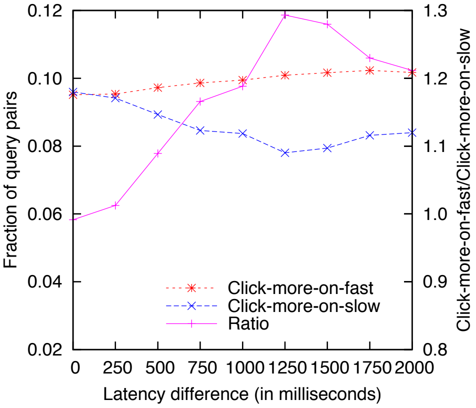


Figure 12: Removing the influence of result quality (by click count).

SIGIR Conf. Research and Development in Information Retrieval, pages 105–114, 2012.

[4] J. D. Brutlag, H. Hutchinson, and M. Stone. User preference and search engine latency. In *In Proc. ASA Joint Statistical Meetings*, 2008.

[5] B. B. Cambazoglu and R. Baeza-Yates. Scalability challenges in web search engines. In M. Melucci and R. Baeza-Yates, editors, *Advanced Topics in Information Retrieval*, volume 33 of *The Information Retrieval Series*, pages 27–50. Springer Berlin Heidelberg, 2011.

[6] J. Dabrowski and E. V. Munson. 40 years of searching for the best computer system response rime. *Interact. Comput.*, 23(5):555–564, 2011.

[7] A. R. Dennis and N. J. Taylor. Information foraging on the Web: The effects of "acceptable" Internet delays on multi-page information search behavior. *Decision Support Systems*, 42(2):810–824, 2006.

[8] J. Gwizdka and I. Lopatovska. The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology*, 60(12):2452 – 2464, 2009.

[9] M. Hassenzahl. Funology. In M. A. Blythe, K. Overbeeke, A. F. Monk, and P. C. Wright, editors, *The Thing and I: Understanding the Relationship Between User and Product*, pages 31–42. Kluwer Academic Publishers, 2004.

[10] J. A. Jacko, A. Sears, and M. S. Borella. The effect of network delay and media on user perceptions of web resources. *Behaviour & Information Technology*, 19(6):427–439, 2000.

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

[12] J. R. Lewis. Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

[13] I. Lopatovska. Searching for good mood: examining relationships between search task and mood. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–13, 2009.

[14] L. McCay-Peet, M. Lalmas, and V. Navalpakkam. On saliency, affect and focused attention. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 541–550, New York, NY, USA, May 2012. ACM.

[15] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, 2008.

[16] F. F.-H. Nah. A study on tolerable waiting time: how long are Web users willing to wait? *Behaviour and Information Technology*, 23(3):153–163, 2004.

[17] H. L. O'Brien and E. G. Toms. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology*, 61(1):50–69, January 2010.

[18] E. Schurman and J. Brutlag. Performance related changes and their user impact. In *Velocity – Web Performance and Operations Conf.*, 2009.

[19] M. D. Smucker. Towards timed predictions of human performance for interactive information retrieval evaluation. In *Proc. Symp. Human-Computer Information Retrieval*, 2009.

[20] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proc. 35th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pages 95–104, 2012.

[21] N. J. Taylor, A. R. Dennis, and J. W. Cummings. Situation normality and the shape of search: The effects of time delays and information presentation on search behavior. *J. Am. Soc. Inf. Sci. Technol.*, 64(5):909–928, 2013.

[22] E. R. Thompson. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of Cross-Cultural Psychology*, 38(2):227–242, 2007.