Stuck in Traffic: How Temporal Delays Affect Search Behaviour

David Maxwell and Leif Azzopardi
School of Computing Science, University of Glasgow
Glasgow, Scotland
d.maxwell.1@research.gla.ac.uk, Leif.Azzopardi@glasgow.ac.uk

ABSTRACT

In this paper we investigate how query response delays and document download delays affect user interactions within a search system. Guided by *Information Foraging Theory* and *Search Economic Theory,* five competing hypotheses relating to the behaviours of searchers in the presence of delays are considered and examined in the context of ad-hoc topic retrieval. A between-subjects laboratory study with 48 undergraduate subjects was conducted. Subjects were randomly assigned to one of four conditions that varied the type of delay experienced. When faced with query response delays, subjects did not examine more documents per query as expected. However, when the total amount of time spent per query (a combination of delay and querying time) increased, subjects did examine more documents per query. When faced with document download delays, subjects did not spent more time within documents. Subjects however did spend longer within documents when subjected to both query and document delays. We found a strong and significant correlation between query time (independent of delay) and the interactions of subjects in terms of the number of queries posed, the number of documents examined, and the depth to which subjects went. These findings contrast with previous works on how delays affect search behaviour, and suggest that the theory needs to be refined to make more credible predictions relating to search behaviours.

Category and Subject Descriptors: H.3.3 **[Information Storage and Retrieval]:** Information Search and Retrieval: Search Process

Terms: Theory, Experimentation, Economics, Human Factors

Keywords: search behaviour, search performance, economic models, interactive information retrieval, query interfaces, query cost

IIiX '14, August 26 - 29 2014, Regensburg, Germany

http://dx.doi.org/10.1145/2637002.2637021

1. INTRODUCTION

Interactive Information Retrieval (IIR) is a non-trivial process where a multitude of factors and variables affect the search behaviours of users [17]. Users interact with search engines in a variety of ways - such as the posing of queries (and any subsequent reformulations), examination of snippets, and the assessment of documents. Performing such interactions requires the user to expend both cognitive and physical effort to some degree, as well as spend time and, in some circumstances, even money [2]. While early research in *Information Retrieval (IR*) considered costs and acknowledged them as central to IR evaluation, little work has explored this issue in detail. However, there has recently been a renewed interest in examining how different costs affect search behaviour [2, 5], and how costs can be included into retrieval measures more explicitly [1, 13, 29, 32]. In such works, cost is often considered as the amount of time spent, and/or the number of interactions performed. In most IIR studies, these variables are reported providing various observations about user search behaviour under different conditions, such as the topic difficulty, searcher expertise, and different interfaces [21]. In this paper, we will focus on the influence of temporal delays on search behaviour, and how being *stuck in web traffic* affects how people interact with search systems and the documents that they link to.

Delays are commonplace on the web. Sometimes, delays can be intermittent and variable, where certain sites take longer to load. This behaviour is a result of a variety of factors, including high volumes of traffic visiting a website, or even intermittent latency issues on the user's browsing device. Delays can also be persistent: for example, searching on a mobile device with limited bandwidth can increase the time required for pages to load, in contrast to viewing the page on a desktop computer with a high-speed Internet connection. These delays have been shown as a source of major frustration to users [7]. Indeed, several studies have examined the effect of page loading delays in the context of browsing the web [15, 25]. It has been shown that the longer a page takes to load, the less favourably users see that page when it finally appears [28]. However, Taylor et al. [33] showed that longer load times resulted in searchers examining more of the downloaded page. Conversely, when query response times increased, it was shown that searchers submitted fewer queries [6, 30]. Clearly, delays have an impact on a user's searching experience, their search behaviours and their information interactions. However, prior research has been rather disjointed, only examining particular aspects of the search process in isolation.

In this paper, we consider two commonly occurring sources of delay: those stemming from the search engine **(query response delays)**, and those stemming from website and network issues **(document download delays)**. The next section summarises the key works that examine different kinds of delay and their effects on user behaviour. In the context of search, these works have focused exclusively on either document download delays when interacting with a list of results [11, 33], or query response delays [6, 30]. This paper examines *both* kinds of delay, and how they influence and affect search behaviour in the context of a complete search task. To frame this research, we then draw upon both *Information Foraging Theory (IFT)* [27] and *Search Economic Theory (SET)* [1] which provide five hypotheses about how a user's search behaviour *s*hould change when faced with delays. These hypotheses - along with our central research question which pertains to the relationship between temporal delays and search behaviour - are outlined in Section 3. We then describe the methodology behind the between-subjects laboratory experiment that was undertaken with 48 undergraduates in Section 4, followed up with our results in Section 5. Finally, we conclude this paper with a discussion and summary of our findings in Section 6.

2. BACKGROUND

Since computer systems have become commonplace in our daily lives, the demands and expectations we place on these systems have increased as the underlying technologies improve. Typically, users expect that an updated piece of software (and/or new hardware configuration) will work *faster* than the previous iteration [9]. It therefore follows that ever faster responses to user requests by computers are a paramount requirement for a given platform [31].

The study of user behaviour in relation to system response times has a long line of research. As early as 1968 Miller [24] suggested the *two second rule*, which states that the majority of interactive system responses should be completed within this period. Users waiting longer than this two second window run the risk of losing the continuity of their thought process, as external influences may disrupt their attention on the task at hand [24, 25]. Miller's two second rule is also in line with a recommendation by Shneiderman [31] and the findings of a study conducted by Nah [25].

While response times in general have been improving, delays on the *World Wide Web (WWW)* are still commonplace. This is because the data requested by users is distributed globally over the Internet. At certain points, network constraints may limit bandwidth. The demand for content is also highly variable (e.g. a higher demand may exist during lunch hours and evenings than during the middle of the night). Consequently, system administrators often spend considerable time and effort trying to improve the responsiveness of the websites and other online services that they provide [3]. Even though Internet connection speeds have increased, multimedia-rich, bandwidth-intensive content (coupled with higher user expectations) means that delays and user frustration are still inevitable [22, 25].

Since the late 1990s, a number of studies have demonstrated that network latency and download speeds can impact how users interact with webpages, as well as their perceived usefulness [10, 15, 18, 25, 28]. These delays are often cited as a complaint by the users of websites [33]. For example, Galletta et al. [15] examined how the behaviour of users changed when faced with page load delays of 0, 2, 4, 6, 8, 10 and 12 seconds. Their results suggest that user performance and behavioural intentions dropped once delays exceeded 2–4 seconds, consistent with Miller's two second rule [24]. After this 'tipping point', users became less favourable to revisiting the given website, and were less likely to recommend the website to others. In addition to these findings, Dennis and Taylor [10] examined delays in the context of interacting with a search results list, and monitored how the behaviour of users changed when faced with a seven second delay. They found that with this longer delay, users spent more time examining the contents of the pages in the search results list. They examined the relationships between the total number of pages viewed by users, the time required for the pages to load, and the volume of information examined on each page. The authors hypothesised that as the page load delay increased, the number of pages examined by users would decrease - but the volume of information examined within each page would increase. Their results provided support for the hypothesis related to page delay time, and the volume of information examined on each page - resulting in *stickier* webpages. A similar finding was also found by Taylor et al. [33]. While these findings are interesting, it should be noted that subjects in the aforementioned studies did not issue their own queries. Instead, they were simply issued with results from a series of predetermined queries. It is therefore unclear whether these findings would hold in the context of a complete, real-world search session.

Page loading delays are now considered by modern search engines when ranking results, such is the perceived importance of this issue [16]. Search engines themselves are now expected to produce the *Search Engine Results Page (SERP)* in a timely manner. A study by Brutlag [6] reports on the effects of time delays on the *Google* search engine. Results from the study showed that the time taken to return search results impacts the number of searches conducted by users. Introducing even a small delay of 400 milliseconds was shown to reduce the number of searches by 0.59 percent over a six week period, using a set of users of the Google search engine. A further study by Schurman and Brutlag [30] - where the impact of server delays was examined by both *Bing* and Google - found similar results with a 500 millisecond delay. A delay of two seconds when returning to the SERP of the Bing search engine was found to reduce the number of queries issued by 1.8 percent. Findings clearly show that user behaviour is much more sensitive to increases in temporal delays when interacting with a search engine. However, Dabrowski and Munson [9] argue that user tolerance for the delay is dependent upon its location and duration. Here, this would mean that users are more tolerant to document download delays than query response delays.

Azzopardi et al. [2] performed a study examining the cost of querying and user behaviour. They devised the *cost-interaction hypothesis*, which states that as the cost of querying increases, the average number of queries will decrease - but the average number of documents examined per query will increase (yielding an increased page stickiness). To test their hypothesis, a search engine with three querying interfaces was implemented, consisting of: ***(1)*** a structured interface (high cost), ***(2)*** a standard querying interface (medium cost), and ***(3)*** a query suggestion interface (low cost). When interacting with the high cost interface, subjects posed significantly fewer queries, spent longer on SERPs, and exam-

ined significantly more documents per query. The subjects of this study were also more likely to consider their efforts as successful when using the high cost interface. These findings were also found by Baskaya et al. [5], who conducted a simulation examining the costs of querying on both a desktop computer (low cost querying) and smartphone (high cost querying). They found that increasing the time to enter a query resulted in a reduction in the number of queries submitted across a variety of querying strategies, and across search sessions of varying lengths. The results implied that as the cost of posing a query increases, the number of queries issued by users decreases.

There is also a line of research in which the key argument is that an increased interaction cost may actually be *beneficial* to users [14]. Studies have shown that people perform better at a given task when an obstacle is in their way [23], and constraints may allow people to clarify and focus on the task at hand [19]. This can already be observed to a certain degree in the cost-interaction hypothesis [2], for example - with increased querying costs leading to stickier results [33].

A further direction on the speed of search was introduced by Kelly [20], who argued that we should be slowing down the user. This concept of *slow search* was considered explicitly in [12, 34]. This movement works on the basis that when system response times are too fast, the number of errors by users increased as they responded to the system too quickly [4]. Teevan et al. [34] and Dörk et al. [12] also state that modern search engines achieve fast response times by sacrificing potential relevance gains in favour of the faster responses. An interesting observation noted by Teevan et al. [34] concerned the abandonment rate of different query types. As SERP load times increased, users posing navigational queries were more likely to abandon their query than those posing informational queries.

3. RESEARCH QUESTION/HYPOTHESES

The driving question behind this research is:

how do delays affect search behaviour?

To supplement this question, we draw upon both IFT [27] and SET [1], and consider what these theories suggest may happen to the behaviours of users as they interact with a search engine.

IFT was used by Dennis and Taylor [10] as a basis to formulate specific hypotheses about how users would interact with a list of search results. Using the *patch model* [26], they interpreted the interaction with the result set as follows. The time spent between documents was referred to as the 'between-patch' time, while the time spent reading and examining documents was called the 'within-patch' time. From IFT, it the follows that an increase in document download time would increase the between-patch time. As a consequence of this, the theory predicts that foragers would spend longer examining the document. This led Dennis and Taylor to the following hypothesis:

> **H1:** *in the face of* ***document download delays,*** *users will spend more time examining information within each document.*

Following a similar analogy, but where the SERP is considered to be the patch, then the between-patch time is the time spent querying, plus any query response delays. It then follows that:

> **H2:** *in the face of* ***query response delays****, users will spend more time on SERPs per query.*

Following on from this is the implication that users would examine more results per query. Taking hypothesis **H1** and **H2** together, we can generate a third, and new, hypothesis:

> **H3:** *when faced with* ***query response delays*** *and* ***document download delays****, users will spend more time examining information with documents and more time examining SERPs per query.*

Conversely, Azzopardi et al. [2] used SET [1] to formulate the *query-cost interaction hypothesis* in the context of a search session. The related hypothesis states that:

> **H4:** *as the relative cost of* ***querying*** *increases, users will pose fewer queries, but examine more documents per query.*

In the study by Azzopardi et al. [2], cost was operationalised as a unit of time. Here, we assume that the relative cost would increase as query response delays increased. However, it should be noted that to increase the relative cost of querying, they introduced an interface which increased the physical effort of entering a query (i.e. more clicks and key presses) which subsequently took a greater amount of time. In this work, we will be examining the influence of adding a query delay, meaning that subjects would expend the same level of effort - but at a greater temporal cost.

Given the analysis by Azzopardi et al. [2], it is also possible to generate the *document-cost interaction hypothesis*, where:

> **H5:** *as the relative cost of* ***assessing a document*** *increases, users will issue more queries and examine fewer documents per query.*

Thus, an increase in download delays would increase the cost of accessing - and therefore assessing - the document. The effect of an increase in document cost is counter to an increase in query cost. Unlike in IFT where the delays reinforce each other, under SET the change in behaviour depends upon the magnitude of the differences in query and document costs. Azzopardi et al. [2] introduced variable β to denote the relative cost of querying to assessing, where:

$$\beta = \frac{c_q}{c_a}$$

and c_q is the cost of a query, with c_a representing cost of assessing a document. It was posited that if the cost of query c_q increased, then β would increase, and this would lead to a change in search behaviour (i.e. the query-cost interaction hypotheses) and similarly - but conversely - for an increase in the cost of a document c_a. In this work, we will be introducing delays, so we can express β as follows:

$$\beta' = \frac{c_q + x_q}{c_a + x_a}$$

where x_q is the query response delay, and x_a is the document download delay. Thus, the time taken to assess and query, along with any additional temporal delays, determines whether β' increases or decreases, and thus whether the number of queries issued increases, decreases or stays the same. If β' increases, then we would expect to see a decrease in the number of queries issued, while the number

of documents examined would increase per query and vice versa if β' decreased. Later in this work, we will use time to represent the costs, and then calculate these β values for each user and condition to provide an indication of what we expect to observe between conditions and users.

4. METHOD

To test the hypotheses stated in Section 3, we conducted a between-subjects laboratory study. The study constituted a search interface where cost was operationalised by introducing additional imposed delays for both querying and loading documents. The developed interface comprised of four conditions, with subjects randomly assigned to either:

1. an interface where systematic delays were not controlled, denoted ***BL*** *(BaseLine);*
2. an interface with an imposed query response delay, denoted ***QD*** *(Query Delay);*
3. an interface with an imposed document download delay, denoted ***DD*** *(Document Delay);* and
4. an interface with imposed query response delays and document download delays, denoted ***QDD*** *(Query and Document Delay).*

For the three conditions utilising them, both query response delays and document download delays were set to five seconds. For query response delays, this meant that an additional five seconds was added onto the initial overhead of waiting for the systematic delay to complete after posing a query. For document download delays, the delay was added after subjects decided which document to click on. A five second delay was guided by the literature as a reasonable value to stimulate a change in user behaviour. Studies have shown that even a small increase in query response delay can impact user behaviour [6, 30], while 'acceptable' delay times for document downloading ranged from three to ten seconds [10], with work by Galletta et al. [15] showing that a document download delay higher than four seconds can lead to a change in behavioural attitudes towards a webpage.

4.1 Corpus, Topics and System

For this experiment, we used the *TREC AQUAINT* test collection. The collection contains over one million newspaper articles from the period 1996-2000. From the *TREC 2005 Robust Track* defined by Voorhees [35], we selected three search topics. The three topics were *piracy* (topic 367), *wildlife extinction* (topic 347) and *curbing population growth* (topic 435). The topics were selected under the belief that they possessed a degree of contemporary relevance, and would be of some interest to our subjects. The three topics also had a similar number of relevant documents (95, 165 and 152 respectively). Topic 367 was used to allow subjects to familiarise themselves with the search interface condition they had been assigned to. This left topics 347 and 435 to be used during the experiment. The ordering of the two topics in which they appeared was rotated using a Latin square.

The developed interface consisted of standard presentational attributes for a search engine, comprised of ten document snippets per page (as shown in Figure 1). The search box provided inline query autocomplete functionality, where suggested terms would be provided based on the user's input. The suggested terms were derived from the collection

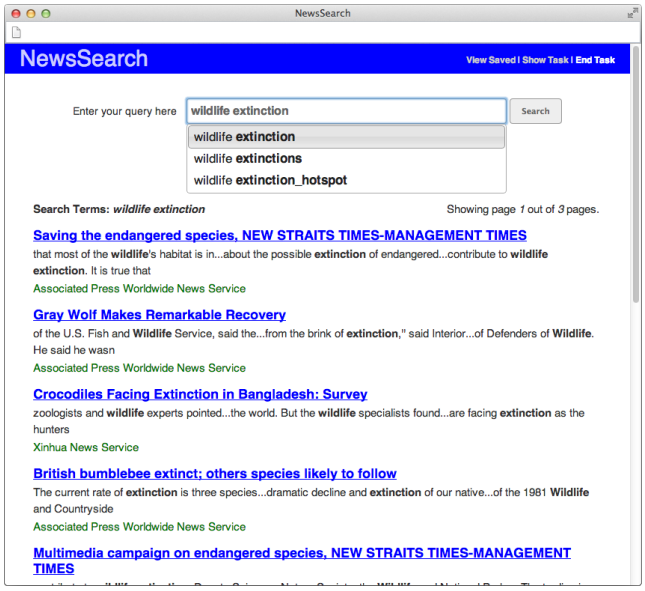

Figure 1: Screenshot of the developed search interface. The image shows a typical SERP that subjects would see, complete with the inline query autocomplete functionality. In the screenshot, results are shown for the query *wildlife extinction.*

used. The query suggestion functionality was included to minimise the number of out-of-vocabulary terms submitted, and to reduce spelling errors in queries submitted to the search engine.

When a query was issued, results would be retrieved by the underlying search engine. For conditions ***QD*** and ***QDD***, the system would then idle - with the interface appearing as unresponsive for five seconds. A *throbber* was placed in the query box for the duration of the query response delay to signify that the system was 'busy'. Conditions ***BL*** and ***DD*** returned the SERP as soon as results were available.

The query response delay was complemented with the document download delay in conditions ***DD*** and ***QDD***. When a subject clicked on a document link in the SERP, a modal box appeared in the middle of the search interface stating that the document was downloading. This persisted for five seconds, before the interface displayed the requested page. During this period, no further interaction could take place. For conditions ***BL*** and ***QD***, documents selected from the SERP were displayed as soon as they were available.

Table 1: Precision figures obtained when querying the developed retrieval system using the titles for the two main TREC 2005 Robust topics.

Topic ***(Topic Number)***	**P@5**	**P@10**	**P@20**
wildlife extinction *(347)*	0.67	0.7	0.67
curbing population growth *(435)*	0.16	0.3	0.2

For the underlying search engine, we used the *Whoosh IR toolkit*[1] with the PL2 retrieval model, where $c = 10.0$. We selected PL2 as it generally performed well on the TREC 2005 Robust topics, and provided different levels of performance over the two main topics chosen. Table 1 provides basic performance metrics for the title queries of the two

[1] `https://pypi.python.org/pypi/Whoosh/`

main topics selected. There was no variation in the underlying search engine used across the four different experimental conditions. Search results were cached on the server in an attempt to minimise systematic query response delays.

For the experiment, two computers were setup with identical monitors, keyboards, and mice. The developed search engine was deployed on a computer on the immediate network to minimise the effects of any network latency issues. Before each subject began the experiment, the web browser history on the computers was cleared by the assistant.

4.2 Logging

To allow us to measure the impact of the cost variations on search behaviour and performance, our search system included a logging component. For each subject involved in the experiment, the logger captured the following key interactions as searching took place:

- the number of **queries issued** - and terms used;
- all the **SERPs viewed**, including which **snippets were hovered over**; and
- all the **documents viewed** by subjects, and the documents **marked relevant**.

From the gathered log data, it was then possible to calculate the time spent performing each activity. For example, we could calculate the time spent issuing queries, examining each SERP, and the time spent by subjects viewing documents. Interactions regarding the practice topic *piracy* were not logged and therefore not included in our analysis.

4.3 Questionnaires

In addition to collecting implicit user actions as mentioned in Section 4.2, we also included three questionnaires for subjects to complete. Below we list the questionnaires included, each with a short description of what data was collected. However, due to space constraints in this paper, we do not report our findings from the listed questionnaires[2].

Demographics Questionnaire. This was included at the very start of each experimental session, where subjects were asked to provide their age, sex, and answer several questions related to their degree.

Pre and Post-Task Questionnaires. Before and after each search task was completed, subjects were asked to complete simple questionnaires. The pre-task questionnaires asked users about their knowledgeability of the search topic, and how easy or difficult they felt it would be to find relevant documents. The post-task questionnaire asked subjects about their experiences of using the search engine, and how good they felt it was at retrieving relevant documents.

4.4 Subjects

A total of 48 undergraduates - 12 per condition - were recruited from the *University of Glasgow*, Scotland, over the period December 2013 to March 2014[3]. Subjects were randomly recruited from a variety of Schools around the University. Of the subjects who were recruited, 32 were male and 16 were female. The mean age of the recruits was 20.3 years ($SD = 2.65$), with a majority in the first year of their studies (41 percent). 75 percent were majoring in a science-based subject, with the remaining 25 percent working towards a humanities-based degree.

[2]Our analysis of these results showed that there was very little difference between responses across the four conditions.

[3]Ethical approval to conduct this experiment was obtained from the University of Glasgow's *College of Science and Engineering Ethics Committee* under reference `ETHICS-CSE01260`.

4.5 Instructions and Incentives

At the start of each session, subjects were briefed by the assistant running the experiment. The briefing consisted of instructing each subject to pretend that they were journalists, and it was their job to use the provided search engine to identify a series of news articles relevant to three topics which they would be provided. As previously mentioned in Section 4.1, the first topic was used to allow the subjects to familiarise themselves with the search interface, with the final two topics constituting the experiment. Subjects were also instructed that approximately 100 documents were identified as relevant by professionals (using TREC relevancy judgements) for each given topic within the searchable collection. They were told to find as many of these relevant documents in the time allotted to them - a maximum of 20 minutes per task. However, if subjects felt that they had collected sufficient documents, or were simply fed up, they could end the task at any time. At the end of the experiment, subjects were presented with totals of the number of documents they correctly identified and the number they incorrectly identified.

Upon completion of the experiment, each subject was compensated for their time with a small reimbursement of £10. Additional funds were awarded to high performing individuals. The performance of each subject considered the number of relevant documents marked, as well as the number of documents marked incorrectly (that were considered as not relevant to the given topic). For those whose performance ranked in the top three for the given topic in their condition, an additional £5 was awarded. With the practice task not being included in these performance calculations, this meant that the maximum *additional* funds a subject could acquire was £10. Providing incentives has been shown to influence the amount of effort people put into their decision making for the task at hand [8]. In this scenario, we wanted to encourage subjects to continue searching and marking documents as accurately as possible. We felt that this was in line with the simulated work task of a journalist, carefully sourcing material for their next story.

5. RESULTS

We present our results across four sections. First, we provide an overview of the experienced delays, before presenting the search behaviours of subjects. For statistical analysis, we used one-way ANOVA followed by Bonferroni's correction post-hoc tests to determine which conditions were significantly different. We then undertook a deeper analysis, inspecting how the search behaviours across the population of 48 subjects were consistent with the hypotheses presented in Section 3. Correlations between the factors in question are measured using Pearson's correlation tests. *For all significance testing reported in this paper, we used an* α *of* 0.05.

5.1 Experienced Delays

Since the focus of this paper is examining the effect of delays on search behaviour, it is important to quantify the actual delays experienced by subjects in each condition. Ta-

ble 2 provides an overview of the delays in each condition with respect to querying and viewing documents. The query delay component consists of two parts: ***(i)*** the system response delay and ***(ii)*** the added delay. For conditions ***QD*** and ***QDD***, the mean query response delay is approximately five seconds, and had a low variance. This is because we added up to five seconds to the wait time based on the system response time. In testing, system response times were optimised and were typically less than two seconds. However, as can be seen from Table 2, the actual experienced delays were typically much higher, where subjects experienced delays of 3.6 and 2.4 seconds per query on average over conditions ***BL*** and ***DD*** respectively. While the ANOVA test showed a significant difference between groups ($F(3, 44) = 8, p = 0.0002$), the follow up test showed that condition ***DD*** had significantly lower query delays than either ***QDD*** and ***QD***. In terms of what subjects experienced on conditions ***BL*** and ***DD***, the delays were more variable and intermittent, depending on whether they issued a query which was in the cache or not. Conversely, subjects in conditions ***QD*** and ***QDD*** experienced a constant and persistent delay of approximately five seconds (with low variability of 0.2 seconds). This should be considered when interpreting our findings. However, on average, delay times in conditions ***QDD*** and ***QD*** were longer.

For document download delays, the approximate delay was five seconds per document across conditions ***DD*** and ***QDD***, with the delay on the other two conditions marginal (typically less than 0.1 seconds). The difference in document delay time over conditions ***DD*** and ***QDD*** was statistically significant, and greater than the other two conditions ($F(3, 44) = 4477, p < 0.05$).

Table 2: Table highlighting the mean delay in seconds (both imposed and systematic) experienced by subjects over each experimental condition. Values enclosed in brackets are standard deviations.

	Condition				
	BL	***QD***	***DD***	***QDD***	**F**
Total Query Response Delay	82.2 (70.4)	115 (54.8)	54.2 (42)	93.1 (53.1)	2.5
Query Response Delay/Query	3.6 (1.8)	4.7 (0.8)	2.4 (1.4)	4.8 (0.9)	8.0
Total Document Response Delay	≈ 0 -	≈ 0 -	282 (120)	199 (42.5)	60.7
Document Response Delay/Doc	≈ 0 -	≈ 0 -	4.8 (0.2)	4.8 (0.2)	4477

5.2 Search Behaviours

Table 3 shows the mean and standard deviation values of the various interactions subjects undertook in each condition. The table reports the number of queries posed, the query length, the number of documents assessed in total, the number of documents assessed per query (and per SERP), and the depth to which subjects viewed documents, as well as their hover depth. We consider the hover depth as the rank of the lowest result on SERPs that subjects hovered over with their mouse and on-screen cursor.

Table 3 also includes how many documents the subjects saved as relevant, how many of these saved documents were deemed to be relevant according to TREC assessors, along with the proportion of saved and relevant divided by the number saved (labelled accuracy). Finally, we present mean precision at 5, 10 and 20 of the SERPs shown to subjects.

Table 3: Mean (and standard deviations) of subject interactions and their search performance on each condition. An asterisk (*) indicates whether there was a statistical difference between conditions.

	Condition				
	BL	***QD***	***DD***	***QDD***	**F**
Number of Queries	22.5 (11.6)	24.7 (11)	23.2 (10.9)	20 (12)	0.4
Query Length	3.58 (0.7)	3.28 (0.9)	3.24 (0.7)	3.7 (0.9)	0.9
Number of Documents Assessed	71.9* (27.6)	65.7 (20.4)	59.3 (25.7)	41.5* (8.8)	4.3*
Documents Assessed/Query	4.2 (2.7)	3.2 (1.6)	3.1 (1.9)	3.2 (2.6)	0.6
Documents Assessed/SERP	2 (1)	1.7 (0.5)	1.4 (0.5)	1.8 (1.4)	0.5
Document Depth/Query	12.9 (6.8)	11.2 (7.9)	12.6 (9.4)	9.6 (5.5)	0.6
Hover Depth/Query	16.5 (7.8)	15.5 (10.6)	18 (11.9)	13.5 (6.3)	0.5
Saved Documents	36.7 (18.1)	37.1 (15)	42.1 (24.5)	24 (8.7)	2.3
Relevant Saved Documents	17.6 (9.2)	13.5 (6.56)	18.8 (12.5)	9.3 (2.6)	3.1
Accuracy	0.5 (0.1)	0.6 (0.2)	0.7 (0.1)	0.6 (0.1)	3.6
P@5	0.32 (0.13)	0.24 (0.09)	0.27 (0.12)	0.24 (0.08)	1.4
P@10	0.3 (0.12)	0.22 (0.08)	0.26 (0.11)	0.23 (0.08)	1.4
P@20	0.24 (0.09)	0.18 (0.07)	0.2 (0.09)	0.19 (0.07)	1.3

From the results of our analysis, subjects issued a similar number of queries across all four experimental conditions - with no significant differences between this finding, and the mean length of queries issued, with $p = 0.79$ and $p = 0.46$ respectively. Interestingly however, subjects in condition ***QD*** issued a higher number of queries than subjects in the other three experimental conditions. Those subjected to delays - especially the document download delay in conditions ***DD*** and ***QDD*** - assessed markedly fewer documents. Indeed, a significant difference existed between conditions ***BL*** and ***QDD*** ($F(3, 44) = 4.3, p = 0.01$). This finding could indicate that **H1** may not hold - but further analysis into the time spent on each document is required - see Section 5.3 for results. We also found no significant difference across the number of documents assessed per SERP viewed - providing evidence against **H2**.

The depths to which subjects went in presented SERPs was also varied across all four conditions. No significant

difference was found for either document assessments ($p = 0.64$) and snippet (hovering) assessments ($p = 0.69$). Subjects in conditions ***BL*** and ***DD*** who were not exposed to a query response delay went to greater depths in the SERP on average. A greater number of snippets were assessed than documents across all four experimental conditions, which was to be expected.

There was no significant difference for both the number of documents marked by subjects ($p = 0.09$) and the number of documents correctly marked as relevant ($p = 0.04$). However, there was a significant difference between conditions ***BL*** and ***QDD*** for the number of relevant documents marked as a ratio of the total number of documents viewed ($F(3, 44) = 3.83, p = 0.02$). The findings suggest that the accuracy of users was affected by the document download delay more than the query response delays imposed on them - and caused subjects to mark more documents relevant on average than when not subjected to the delay. However, the compounding effect of both query response and document download delays actually decreased the total number of documents marked, and thus reduced the number of correctly identified documents.

Across all query precision readings, no significant differences were observed across the four experimental conditions. This demonstrates that the performance of queries may not be significantly impacted by query response delays. P@1 for condition ***BL*** was observed at 0.3286, almost 0.1 greater than the P@1 values recorded for the other three conditions. This provides evidence that the queries posed by subjects in condition ***BL*** were on average more likely to return a relevant document in the first position of the corresponding SERP. That is, subjects in condition ***BL*** posed more effective queries than those in other conditions.

5.3 Time Spent Searching

Table 4 reports statistics regarding the time subjects spent engaged in the various aspects of the search process across both topics undertaken by subjects. We report the mean time spent formulating a query (excluding any imposed delay), the mean time spent by subjects on each SERP page viewed, and the mean SERP time per query. Table 4 also shows the mean for the time spent on each document, and the average session time.

While there were no significant differences between the times to perform the various interactions, it is interesting to note a number of trends. Subjects in condition ***QDD*** spent longer on average examining SERPs and formulating their queries than subjects in the other three experimental conditions. This also meant that for each query issued, those in condition ***QDD*** spent longer on SERPs per query than subjects in the other three conditions. The standard deviation is however quite large, suggesting a large time spread for individual subjects. This is in contrast to findings presented in Section 5.2 which showed no significant difference in the number of documents assessed per SERP viewed. This therefore suggests that when faced with both query response and document download delays, subjects spent longer examining snippets on SERPs, and that they were more conservative in which documents they examined further - which is at odds with hypothesis **H3**. However, the data also shows that subjects in condition ***QDD*** on average spent longer on each document viewed than subjects in the other three conditions, which in turn provides some support for **H3**.

Table 4: Mean times (in seconds, with (SDs)) recorded for various interactions across the four experimental conditions.

	Condition				
	BL	***QD***	***DD***	***QDD***	**F**
Time/Query	7.7 (1.7)	8 (1.7)	9.6 (5.7)	9.9 (3.1)	1.2
Time/SERP	19.7 (9.4)	18.9 (6.1)	22.8 (7.2)	30.6 (17.9)	2.8
SERP/Query	39.7 (24)	37.3 (20.7)	49.2 (28.1)	53.3 (35)	0.9
Time/Document	19.7 (9.5)	18.8 (7.43)	19.5 (11.5)	27.8 (9.32)	2.4
Average Session Time	1188 (19.2)	1192 (17.5)	1186 (12.8)	1174 (21.1)	0.2

5.4 Correlation Analysis

Thus far, we have seen little evidence to support the hypotheses which we presented in Section 3. In this section, rather than comparing across conditions, we decided to consider the hypotheses with respect to all 48 subjects. We took this approach as different subjects spent varying amounts of time querying and assessing, and were also subjected to delays of varying lengths.

H1 ***(DD delays lead to more time being spent within each document).*** Earlier in Section 5.2, we showed that findings in Table 3 provided mixed evidence to suggest that subjects increased the time spent on each document when faced with document download delays. To examine this on an individual basis, we performed a correlation analysis between the document download delay and time spent on documents. However, this yielded only a low, non-significant correlation ($r = 0.23$, $p = 0.12$). This follow up analysis indicates that in the context of searching and browsing, an increase in document download time does not necessarily result in spending more time within a document. This finding is in contrast to results presented by Dennis and Taylor [10], where they did find a relationship between document download times and the time spent within a document. However, no querying was performed by subjects in that study - only a predefined results list was examined. Furthermore, the imposed delay on their subjects was greater, at seven seconds.

H2 ***(QRDs result in more time spent on each SERP).*** To determine if there was a relationship between the query response delay and the amount of time spent on SERPs, we again looked at the correlations between these variables. A low, non-significant correlation was observed ($r = 0.2$, $p = 0.17$). We speculated that the time spent querying should also be included, as it contributes to the between patch time as well. This led to a slightly stronger (but still low correlation) of $r = 0.25$, which approached borderline significance ($p = 0.084$). Interested by this result, we then included the time subjects spent examining documents - or within patch time. We found a stronger, significant correlation ($r = 0.46$, $p = 0.001$). This finding suggests that an increase in querying time has an effect not only on the time spent on SERPs, but on documents, too. As such, **H2** should be altered to include document examination time.

H3 ***(QRDs and DD delays result in more time spent in documents and on SERPs).*** Here, we examined if the sum of query and document delays experienced correlated with users spending more time examining documents, and more time examining SERPs. Here, we found stronger correlations which suggest that both factors were coming into play. With respect to time spent within a document, the correlation was $r = 0.26$ ($p = 0.07$). However, a significant correlation was observed with respect to SERP time ($r = 0.39$, $p = 0.01$). Encouraged by this result, we then hypothesised that the perceived impact of imposed query response delays could be influenced by the time subjects spent querying (e.g. a relatively small time period spent querying would mean the imposed query response delay felt longer). When incorporating query time as well, we found stronger and significant correlations between Query Time ***(QT)*** + Query Response Delay ***(QRD)*** + Document Download Delay ***(DDD)*** versus the Document Time (the time spent on a document) ***(DT)***, the correlation was $r = 0.50$ ($p < 0.01$), and versus time spent on the SERP page ***(SERPT)*** the correlation was $r = 0.35$ ($p = 0.01$). This suggests that both delays and the *total* query time must be considered, and that they do have a bearing on search behaviours. These findings provide support for **H3**. This also suggests that if we want to show evidence for **H1** and **H2**, then we need an experiment that is far more controlled. This would then allow us to isolate and change the variable of interest - something which is of course difficult to achieve in practice.

H4 ***(fewer queries posed, and more documents examined when querying costs increase)*** **and H5** ***(fewer documents examined and more queries issued when document assessment costs increase).*** It should be noted that **H4** and **H5** are dependent on β, which represents the relative cost of querying to assessing. In Section 3, we defined β and β' (which included delays). Here, we computed these values for each subject, and then examined whether they were correlated with the number of queries subjects submitted (r_q), the number of documents they examined per query (r_d), the depth they went to per query (r_{dep}), their search accuracy (r_{acc}) and the number of documents they saved which were TREC relevant (r_{rel}). Table 5 provides a summary of the correlations for these variables and the β values. We also examined different time factors ***QT***, ***QRD***, ***DT***, ***DDD*** and ***SERPT***.

If we first examine the correlations with respect to β without considering delays, we note that there are only weak correlations with the number of queries issued and the number of documents examined per query. However, when we include delays (i.e. β'), we then begin to observe a significant correlation between β' and the depth per query. These results seem to indicate that the SET hypotheses **H4** and **H5** do not hold. However, Azzopardi et al. [2] suggested that the time spent on the SERP page should also be accounted for. Consequently, we included the SERP time to denote the relative cost as:

$$\beta'' = \frac{c_q + x_q + s_q}{c_a + x_a}$$

where s_q is the time spent per query page. Using this formulation of the relative cost, we obtained significant and strong correlations where we observed that as β'' increased, fewer queries were posed ($r = -0.4$), while more documents were examined per query ($r = 0.7$) and that subjects went to greater depths ($r = 0.8$). With this modification, much like the modification required to contextualise the IFT results, we find evidence to support **H4** and **H5** given β''. Figure 2 shows a plot of the β values and number of queries issued (top, in red) and the number of documents examined per query (bottom, in blue).

Table 5: Correlations between factors and interactions. An asterisk (*) denotes that the Pearson's correlation was significant.

Factor	r_q	r_d	r_{dep}	r_{acc}	r_{rels}
QRD	−0.05	0.04	−0.08	−0.08	−0.20
QRD+QT	−0.18	−0.23	−0.26	0.09	−0.41*
QRD+QT+SERPT	−0.82*	0.77	0.68*	−0.30*	−0.23*
DDD	−0.11	−0.09	−0.12	0.02	−0.09
DDD+DT	−0.32*	−0.23	−0.35*	0.09	−0.58*
DDD+QRD	−0.11	−0.09	−0.12	−0.04	−0.19
DDD+QRD+QT+DT	−0.31*	−0.26	−0.36*	0.1	−0.60*
β	0.17*	0.07	0.27*	−0.03	0.56*
β'	0.15*	0.17	0.33*	−0.08	0.40*
β''	−0.4*	0.70*	0.80*	0.16	0.66*

In Table 5, we have also included the correlation between the different variables and the performance of subjects in terms of the number of documents they marked as TREC relevant, along with their search accuracy. Of note, there is a low negative correlation between the number of TREC relevant documents subjects found, and the query delay ($r = -0.2$). However, when taken together with the time spent querying, the correlation is moderate but significant ($r = -0.41$), suggesting that as the delays and the amount of time spent querying increased, the number of relevant documents found decreased. Similarly, as the amount of time spent assessing a document and the document delay increased, there was also a moderate negative and significant correlation with the number of relevant documents found ($r = -0.58$). This result intuitively makes more sense, because the more time you spend reading documents, the fewer documents you can examine for relevance. However, the precision of the subjects did not seem to be affected by the time taken to undertake such actions.

6. DISCUSSION AND FUTURE WORK

In this paper, we investigated how query and document delays affect search behaviour. To provide a theoretical underpinning to this investigation, we used IFT and SET to provide hypotheses about search behaviours when temporal delays were present. To this end, we conducted a laboratory study with 48 subjects who were allocated to one of four conditions where differing delays were imposed. We found mixed support for the hypotheses stemming from IFT theory (**H1**-**H3**), but strong evidence for **H3**, once we revised it to also consider the query time as well. Here, we observed that subjects spent longer in documents and on SERPs as the sum of the query delay, document delays and query time increased. We also found strong evidence to support the hypotheses from SET (**H4** and **H5**). Specifically, we observed that as the relative cost of querying increased, then subjects posed fewer queries and examined more documents per

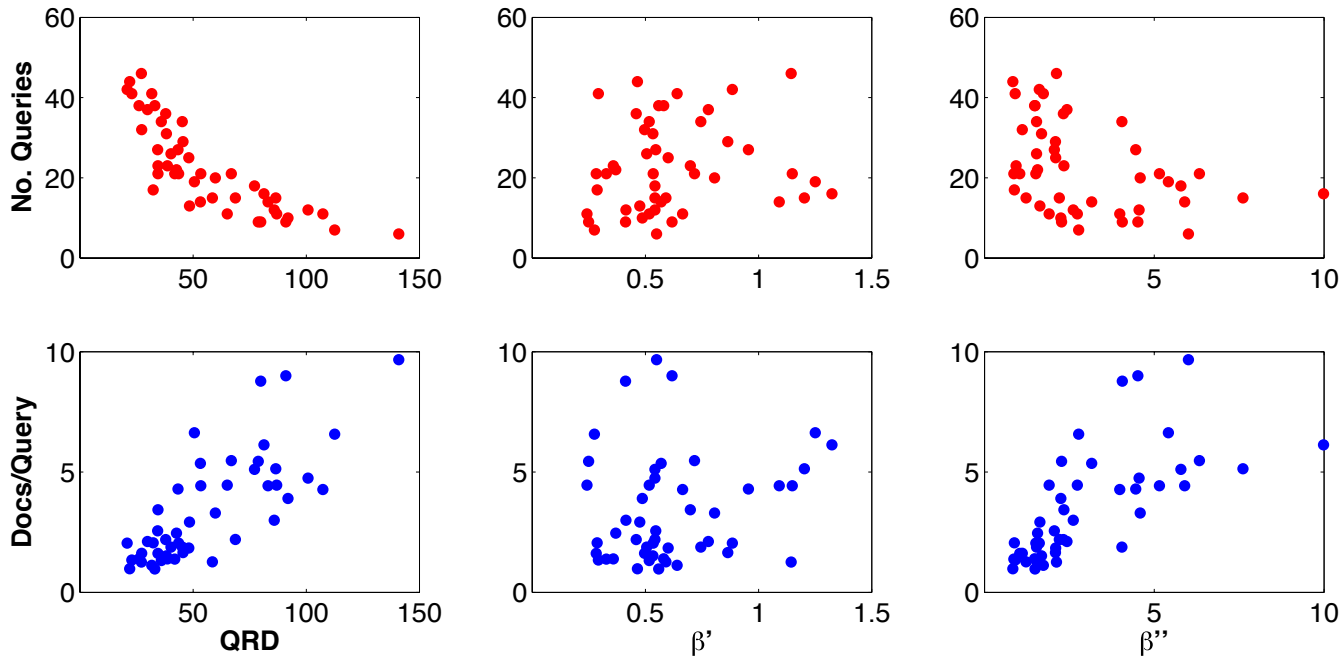

Figure 2: A series of plots showing the correlations between various factors. Along the top are plots of the number of queries posed by each subject. Along the bottom are plots of the number of documents posed per query. From left to right, each of these are correlated with factors *QRD*, β' and β''.

query. However, it was only when we considered the total time per query (delays, query time and the time spent on each SERP) did we find that this relationship was stronger and significant. This suggests that SET needs to be revised to account for interactions on the SERP. This finding also provides support for the observations made by Azzopardi et al. [2] in that a revision of SET is necessary.

An interesting finding was that the query delay is not correlated with querying and assessing search behaviour, whereas the time spent issuing queries was moderately correlated. This suggests that the effort involved in querying - as denoted by the time spent issuing the queries - affects search behaviour more so than the delay time. The delay does not require the subject to expend more effort, just simply to wait. From an economic perspective, this wait time is similar to a fixed cost that has to be paid regardless, and so will not affect the interaction - unless the cost is of course so high that users would be unwilling to pay it. Instead, it is the variable cost of query time that is more important. This is due to the fact that the user has to decide on how much more effort and time they should invest in crafting a quality query, versus how much more they will get out of posing a better query. Given the negligible effect of the delays, this finding motivates a deeper study into the difference between effort and time, and how to quantify the cost of searching and interacting.

A limitation of the study is that we did not control the query system response delay sufficiently. As such, our findings show the difference between variable query delays and more constant, fixed delays. Nonetheless, we have examined two kinds of delay in the context of the wider search process, and not in isolation like in previous works. In future work, we will examine how users behave when there are no delays. In addition, examining behavioural changes with imposed delays *greater* than five seconds would for example provide evidence to determine if the 2–4 second behavioural 'tipping point' highlighted by Galletta et al. [15] holds. Future studies should also examine if the behaviours described in this paper generalise across a wider range of topics, or if they hold when subjects provide their own information needs. Furthermore, our interpretation of IFT is also based on that of Dennis and Taylor [10]. As there are different ways to interpret the patch model, it would be interesting to explore different interpretations of the search process. Finally, while delays are less likely to be experienced in a desktop setting, it would be interesting to explore how delays affect search interactions when using devices such as mobile phones. These devices generally have connection speeds that are slow, and thus users may experience high latency.

Invariably, it seems that being stuck in traffic is not desirable and affects search performance in terms of the number of relevant documents identified. However, in this study it seems to have had little impact on search behaviour, or the accuracy of the judgements made by subjects. Instead, we found that the time spent querying, examining the SERP and assessing documents play a larger role in shaping the search behaviours of subjects.

Acknowledgments: We would like to express our thanks to Horatiu Bota, Simon Jouët, Sean McKeown and Kyle White for their feedback and assistance on this work. We would also like to thank the 48 University of Glasgow undergraduates who took part in this study.

References

[1] L. Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th ACM conference on re-*

search and development in information retrieval (SIGIR), pages 15–24, 2011.

[2] L. Azzopardi, D. Kelly, and K. Brennan. How query cost affects search behavior. In *Proceedings of the 36th ACM conference on research and development in information retrieval (SIGIR)*, pages 23–32, 2013.

[3] R. Baeza-Yates, V. Murdock, and C. Hauff. Efficiency trade-offs in two-tier web search systems. In *Proceedings of the 32nd ACM conference on research and development in information retrieval (SIGIR)*, pages 163–170, 2009.

[4] R. E. Barber and H. C. Lucas, Jr. System response time operator productivity, and job satisfaction. *Communications of the ACM*, 26(11):972–986, 1983.

[5] F. Baskaya, H. Keskustalo, and K. Järvelin. Time drives interaction: simulating sessions in diverse searching environments. In *Proceedings of the 35th ACM conference on research and development in information retrieval (SIGIR)*, pages 105–114, 2012.

[6] J. Brutlag. Speed matters for google web search. `http://goo.gl/t7qGN8` (retrieved on March 26th, 2014), 2009.

[7] I. Ceaparu, J. Lazar, K. Bessiere, J. Robinson, and B. Shneiderman. Determining causes and severity of end-user frustration. *International Journal of Human-Computer Interaction*, 2004.

[8] E. H. Creyer, J. R. Bettman, and J. W. Payne. The impact of accuracy and effort feedback and goals on adaptive decision behavior. *Journal of Behavioral Decision Making*, 3(1):1–16, 1990.

[9] J. Dabrowski and E. V. Munson. 40 years of searching for the best computer system response time. *Interacting with Computers*, 23:555–564, 2011.

[10] 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*, 42:810–824, 2006.

[11] S. Dennis, P. Bruza, and R. McArthur. Web searching: a process-oriented experimental study of three interactive search paradigms. *Journal of the American Society for Information Science and Technology*, 53(2):120–133, 2002.

[12] M. Dörk, P. Bennett, and R. Davies. Taking our sweet time to search. In *Proceedings of CHI 2013 Workshop on Changing Perspectives of Time in HCI*, 2013.

[13] N. Fuhr. A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3):251–265, 2008.

[14] K. Fujikawa, H. Joho, and S. Nakayama. Constraint can affect human perception, behaviour, and performance of search. In *Proceedings of the 14th International Conference on Asia-Pacific Digital Libraries (ICADL)*, pages 39–48, 2012.

[15] D. F. Galletta, R. Henry, S. McCoy, and P. Polak. Web site delays: How tolerant are users? *Journal of the Association for Information Systems*, 5:1–28, 2003.

[16] Google Webmaster Central Blog. Using site speed in web search ranking. `http://goo.gl/EM87T` (retrieved on March 28th, 2014), 2010.

[17] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., 2005.

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

[19] R. Junco and S. R. Cotten. Perceived academic effects of instant messaging use. *Computers & Education*, 56(2):370–378, Feb. 2011.

[20] D. Kelly. Cognitive consequences of search. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX '12, pages 2–2, 2012.

[21] D. Kelly and C. Sugimoto. A systematic review of interactive information retrieval evaluation studies, 1967-2006. *Journal of the American Society for Information Science and Tech.*, 64(4):745–770, 2013.

[22] T. Leighton. Improving performance on the internet. *Communications of the ACM*, 52(2):44–51, 2009.

[23] J. Marguc, J. Förster, and G. A. V. Kleef. Stepping back to see the big picture: When obstacles elicit global processing. *Journal of Personality and Social Psychology*, 101(5):883–901, 2011.

[24] R. B. Miller. Response time in man-computer conversational transactions. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*, AFIPS '68 (Fall, part I), pages 267–277, 1968.

[25] 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.

[26] P. Pirolli and S. Card. Information foraging in information access environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '95, pages 51–58, 1995.

[27] P. Pirolli and S. Card. Information foraging. *Psychological Review*, 106:643–675, 1999.

[28] J. Ramsay, A. Barbesi, and J. Preece. A psychological investigation of long retrieval times on the world wide web. *Interacting with Computers*, 10(1):77 – 86, 1998.

[29] T. Sakai and Z. Dou. Summaries, ranked retrieval and sessions: A unified framework for information access evaluation. In *Proceedings of the 36th ACM conference on research and development in information retrieval (SIGIR)*, pages 473–482, 2013.

[30] E. Schurman and J. Brutlag. Performance related changes and their user impact. `http://goo.gl/Tnb5c` (retrieved on March 28th, 2014), 2009.

[31] B. Shneiderman. Response time and display rate in human performance with computers. *ACM Computing Surveys*, 16 (3):265–285, 1984.

[32] M. D. Smucker and C. L. Clarke. Time-based calibration of effectiveness measures. In *Proceedings of the 35th ACM conference on research and development in information retrieval (SIGIR)*, pages 95–104, 2012.

[33] 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. *Journal of the American Society for Information Science and Tech.*, 64 (5):909–928, 2013.

[34] J. Teevan, K. Collins-Thompson, R. W. White, S. T. Dumais, and Y. Kim. Slow search: Information retrieval without time constraints. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval*, HCIR '13, pages 1:1–1:10, 2013.

[35] E. M. Voorhees. Overview of the trec 2005 robust retrieval track. In *Proceedings of TREC-14*, 2006.