How Query Cost Affects Search Behavior

Leif Azzopardi
School of Computing Science, University of Glasgow
Glasgow, United Kingdom
leif@dcs.gla.ac.uk

Diane Kelly
School of Information and Library Science, University of North Carolina
Chapel Hill, NC, USA
dianek@email.unc.edu

Kathy Brennan
School of Information and Library Science, University of North Carolina
Chapel Hill, NC, USA
knb11@live.unc.edu

ABSTRACT

We investigate how the cost associated with querying in the context of information retrieval affects how users interact with a search system. Microeconomic theory is used to generate the *cost-interaction hypothesis* that states as the cost of querying increases, users will pose fewer queries and examine more documents per query. A between-subjects laboratory study with 36 undergraduate subjects was conducted, where subjects were randomly assigned to use one of three search interfaces that varied according to the amount of physical cost required to query: Structured (high cost), Standard (medium cost) and Query Suggestion (low cost). Results show that subjects who used the Structured interface submitted significantly fewer queries, spent more time on search results pages, examined significantly more documents per query, and went to greater depths in the search results list. Results also showed that these subjects spent longer generating their initial queries, saved more relevant documents and rated their queries as more successful. These findings have implications for the usefulness of microeconomic theory as a way to model and explain search interaction, as well as for the design of query facilities.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval:Search Process; H.3.4 [**Information Storage and Retrieval**]: Systems and Software:Performance Evaluation

General Terms

Theory, Experimentation, Economics, Human Factors

Keywords

Search Behavior, Economic Models, Production Theory, Interactive Information Retrieval, Query Interfaces, Query Cost

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

1. INTRODUCTION

During interactive information retrieval (IIR), users perform various interactions, such as posing queries, examining snippets and evaluating documents. Each action requires a certain amount of effort and comes at a cost, whether mental, physical, temporal, fiscal, or a combination thereof. In the early days of search system research, cost and utility figured prominently in both systems-centered and user-centered evaluation frameworks [6, 36]; however, less attention has been devoted to costs in contemporary information search research. Fiscal costs are presumed to be less important because search tools and information are freely available online. Mental and physical costs have been incorporated into evaluation measures as fixed parameters (for example, by introducing discounting based on rank [24]) and used to characterize interactions [25], but they have rarely been studied as independent variables because they are difficult to manipulate and measure. While many interactive search studies incorporate effort-based measures, such as the number of documents examined, number of queries issued and amount of time spent performing different actions [35, 25], few studies have attempted to model how interaction costs shape search behavior [27]. There have been some exceptions, though, where information seeking and retrieval behavior has been formally modeled using a cost-benefit framework [2, 31, 34].

In this work, we explore how the costs associated with querying affect search behaviors in the context of IIR. This is an important question to address because understanding the relationship between cost, behavior and performance might help explain how and why users interact with search systems in particular ways, and subsequently, enable designers of such systems to influence user interaction and search behavior. We ground our experiment using the recently proposed Search Economic Theory [2], which uses microeconomics to model the search process and provides a means to reason about interaction, cost and performance. Using this theory, we formulated the *cost-interaction hypothesis* that states: as the cost of querying increases, users will pose fewer queries and examine more documents per query. We conducted a laboratory experiment with 36 subjects and three interfaces that varied according to query cost, where cost was operationalized as time required to submit a query. Our findings show that subjects who used the more costly query interface submitted significantly fewer queries and examined significantly more documents per query than subjects who used the interfaces with lower querying cost.

2. BACKGROUND

Cost has had a long history in IR research and has figured prominently in both systems-centered and user-centered research. Cost have been defined in a number of ways, including as mental, physical, temporal and fiscal cost. In the pre-Web era, research often focused on fiscal costs since search tools and online information were not free; many measures of fiscal costs focused on the amount users were willing to pay for search results [6, 9, 41]. For example, Cooper [10] proposed that users associate dollar amounts with search results as a way to understand subjective utility and Salton [36] proposed a number of cost-based measures related to the operational environment and response times. Today, only a few IR measures attempt to incorporate some type of cost into their computation, including nDCG [24], RBP [28], and time-biased gain [40].

While fiscal costs are no longer investigated as much, temporal cost, and more specifically response time, continues to be an important variable in research [12]. Since the late 1990s, a number of studies have demonstrated that network latency and download speeds impact how people interact with Web pages and their evaluations of information and website quality (e.g., [14, 23, 29]). In a recent study of time delays and search behavior, Taylor et al. [42] investigated the relationship between the amount of time Web pages take to load, the number of pages people viewed and the amount of information examined per page. Taylor et al. hypothesized that as response time increased, the number of pages searched would decrease and the amount of information examined per page would increase. The researchers hypothesized a step relationship between response time and search behavior rather than a linear or curvilinear relationship (that is, changes in behavior would only occur after a critical time delay point). Results showed partial support for the hypothesis related to response time and number of pages examined, and full support for the hypothesis related to response time and the amount of information examined per page. While these results provide evidence that cost, as measured by response time, impacts search behavior, the task explored in this study presented participants with a set of static Web pages, and query cost and participants' interactions with result pages were not examined.

There is some evidence that introducing time delays and query constraints can impact search behavior. Brutlag [7] reports in a blog about research conducted at Google that the time taken to return search results impacts the number of searches conducted by users. An increase in 400 milliseconds was shown to reduce the number of searches by 0.59% across a six week period on the Google search engine. Fujikawa et al. [19] constrained the number of queries a user could issue, ostensively making queries more valuable, and found that participants whose querying was constrained posed fewer queries and examined more documents per query, while participants who were not constrained submitted more queries and examined fewer documents per query. While this study was not focused on query cost, the results suggest that limitations on people's abilities to input queries can impact search behavior. Several studies have examined how query input facilities impact the types and properties of queries entered by users, and how this subsequently impacts search outcomes [1, 5, 17], but these relationships were not investigated in a cost-benefit framework. Furthermore, these studies have mainly focused on encouraging different querying behaviors, rather than understanding how the query facility and the cost of querying shapes the entire search interaction.

Recently, Baskaya et al. [3] used a simulation to study the cost of querying on two different devices. Query cost was measured as time and set to different speeds related to how long it takes the average person to enter a query using a desktop computer and smart phone. Baskaya et al. 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 sessions of different lengths. Essentially, this simulation suggests that as query cost increases, the number of queries issued will decrease. However, the findings have yet to be empirically validated.

Researchers have also used physical costs, or the amount of effort a person exerts during search, as a way to evaluate IR systems and user interaction. These effort-based measures include number of queries issued, number of result pages evaluated and documents viewed [27]. However, few laboratory studies have attempted to model behavior or understand if and how different costs shape behavior, or how different interfaces are associated with different costs. Studies using large scale search log data have made some progress on modeling effort and search behavior (e.g., [16, 46]), but since physical cost cannot always be manipulated in operational environments, these studies present only a partial view of search behavior under particular circumstances. In addition, while these studies indicate how people act with a given technology, they do not show how people might act given different interface costs.

Researchers have also tried to model search costs by examining mental effort. This represents a much smaller body of work (but no less important, since searching is, by nature, a mental activity) presumably because of the difficulty of measuring mental effort. Both Dennis et al. [15] and Gwizdka [20] have examined the cognitive costs of different interfaces using cognitive load theory. In these works, the authors conducted experiments to estimate the mental effort required to undertake various search interactions during the search process. However, the relationship among cost, performance and interaction was not explored.

In terms of theoretical research there has been a number of proposals that model costs and search behavior within a cost-benefit framework [2, 18, 4, 31, 32, 33, 34]. Bates [4], for example, suggests one of the search tactics users adopt while interacting with a system is to make decisions about whether to pursue the current strategy or to change strategy, depending on a cost-benefit analysis. While Bates did not formally pursue this idea, Russell et al. explored this in their work on the cost of sense-making [34]. Here they analyzed the possible actions a user could take during the information seeking process in terms of the gain that would be accumulated over time. Then actions could be compared under the assumption that users would try to maximize gain while minimizing total cost. In this process, gain can be seen as the amount of relevant information found (or the value of the relevant information found). Information Foraging Theory (IFT) [31] provides a theoretical grounding for these ideas. IFT models how users would act and behave within heterogeneous information environments in ecologically valid ways. Specifically, it proposes that information seekers aim to minimize effort and maximize gain as they move between information patches, follow scents and assume a particular information diet. In experiments using a clustering inter-

face based on the Scatter-Gather principle, it was shown that users tend to act in an ecologically valid manner (that is, they conserve effort while seeking the most gain) [33].

Another way to formally model the information seeking process is through microeconomics. Azzopardi [2] proposes Search Economic Theory (SET) as a way to predict and explain search behavior. The model consists of a gain function and a cost function, which are parameterized by the type and number of interactions performed during the search process. Like the other formal models, the model assumes people will seek to minimize costs and maximize gain. Unlike the previous work, the theory was specifically developed to model the interaction between a user and an information retrieval system. As a result, the theory may provide useful insights into search behavior and can guide empirical investigation. In this paper, we explore how query costs affect search behavior under SET and then describe an empirical study that investigates this theory.

3. SEARCH ECONOMIC THEORY

In Azzopardi [2], the search process was modeled using an analogy to Production Theory [44]. In Production Theory, a firm takes inputs (i.e., capital and labor) and converts them to output (i.e., widgets). When applied to search, a user with a search engine is considered the firm, the user's interactions are considered inputs and the relevant information found is considered as the output (and measured by Cumulative Gain (**CG**) [24]). Azzopardi defined the inputs as the number of queries posed (**Q**) and the number of documents assessed per query (**A**). A functional relationship was proposed such that performance was related to interaction as follows: $\mathbf{CG} = g(\mathbf{Q}, \mathbf{A})$. This function (referred to as the search production function) denoted the upper bound on output given a specific input combination (i.e., the maximum that can be produced with the given inputs). It was shown that for several standard information retrieval models (such as Boolean, VSM with TFIDF and BM25 [11]) the function $g(\mathbf{Q}, \mathbf{A})$ could be modeled closely with a Cobbs-Douglas production function [44]:

$$g(\mathbf{Q}, \mathbf{A}) = \mathbf{k}.\mathbf{Q}^{\mathbf{b}}.\mathbf{A}^{(1-\mathbf{b})} \quad (1)$$

where **k** denoted how well the hypothetical user could convert actions into relevant documents using the search system, and **b** was a mixing parameter which regulated the interplay between querying and assessing. The following cost model $c(\mathbf{Q}, \mathbf{A})$ was then used to ascribe a total cost to the interactions undertaken [2]:

$$c(\mathbf{Q}, \mathbf{A}) = \mathbf{a}.\mathbf{Q} + \mathbf{Q}.\mathbf{A} \quad (2)$$

where **a** denoted the relative cost of querying to assessing. The cost of assessing documents was assumed to be 1 (where the total number of documents assessed was **Q** multiplied by **A**).

Given this model of interaction defined by the gain and cost functions, it is possible to explore how the search behavior of a user would change when different variables are manipulated. We conducted a simulation to illustrate what changes in search behavior we could expect to see when query cost increases under the proposed model. We explored a range of relative querying costs ($\mathbf{a} = 0.5, 1, 2, 4$) across various search production functions where **b** was varied from 0.5 to 0.6 and **k** was set to 3. Then we set $\mathbf{CG} = g(\mathbf{Q}, \mathbf{A}) = 30$

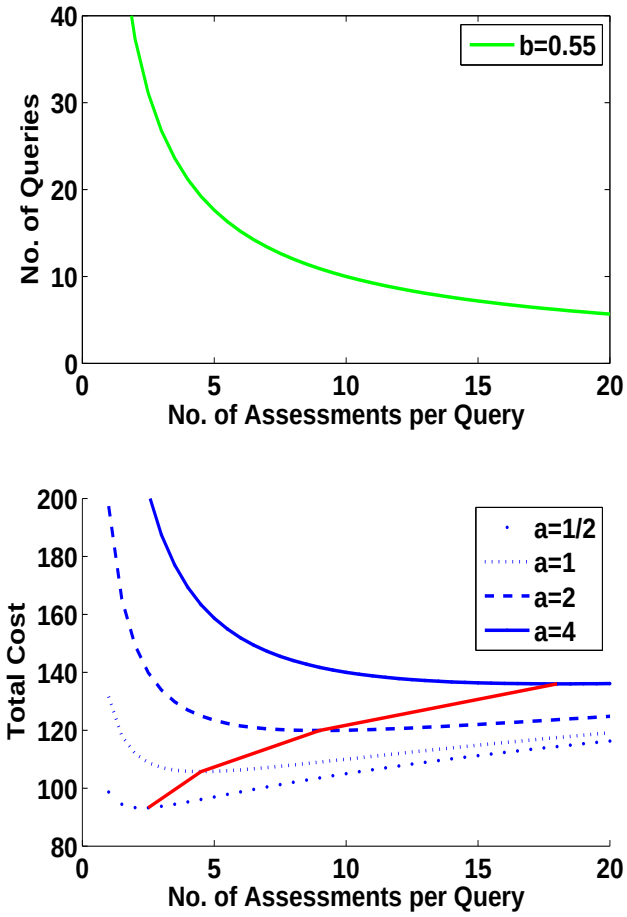

Figure 1: An example production function when b=0.55 (top), and the corresponding cost curves for different levels of a (bottom). The red line indicates the minimum cost point on each cost curve.

which models the situation where users are looking for 30 relevant documents (assuming binary relevance for computing cumulative gain)[1].

Figure 1 shows the search production function when $\mathbf{b} = 0.55$ (top plot) and the corresponding cost curves (bottom plot). Each point on the production curve represents the number of queries (**Q**) and the number of assessments per query (**A**) required to find 30 relevant documents. When $\mathbf{b} = 0.55$, there exists a number of possible combinations of inputs that would yield the desired output: a user could, for example, issue approximately 10 queries and assess 10 documents per query, or issue approximately 4 queries and assess 20 documents per query to obtain the same gain.

The bottom plot shows the corresponding cost curves associated with the gain function when $\mathbf{b} = 0.55$. The red line indicates the minimum cost point on each cost curve (i.e., the combination of inputs which minimizes the cost given the relative cost of querying). It is clear from this plot that as the relative query cost increases, the number of assessments per query also increases (and this results in a corresponding decrease in the number of queries issued). This trend

[1]Note that when we varied **k** this only affected the number of queries required not the number of assessments per query. So as **k** increased, **Q** decreased.

		Query Cost			
	a	**0.5**	**1.0**	**2.0**	**4.0**
b=0.51	**A**	12.5	24.5	49.0	98.0
	Q	8.1	4.2	2.2	1.1
	mc	104.9	107.8	110.8	113.8
b=0.53	**A**	4.0	8.0	15.5	31.5
	Q	22.5	12.2	6.8	3.6
	mc	101.4	109.7	118.6	128.3
b=0.55	**A**	2.5	4.5	9.0	18.0
	Q	31.1	19.2	10.9	6.2
	mc	93.3	105.7	119.9	136.0
b=0.57	**A**	1.5	3.0	6.0	12.5
	Q	41.8	24.8	14.7	8.5
	mc	83.7	99.2	117.6	139.4
b=0.59	**A**	1.0	2.5	4.5	9.0
	Q	49.5	26.2	17.4	10.8
	mc	74.3	91.7	113.2	139.9

Table 1: For different production curves characterized by b, the A and Q that result in the minimum cost (mc) for the relative query cost a is shown. As a increases, Q decreases and A increases.

is similar for different values of **b**. Table 1 shows a number of outcomes where **b** is varied between 0.51 and 0.59. The combination of **Q** and **A** that minimizes the total cost for each **a** is shown for each gain function with value **b**. These results show more generally that the model predicts the following: as relative query cost increases (i.e. $\mathbf{a} = 0.5 \rightarrow 4$), to minimize overall cost, a user should decrease the number of queries issued and increase the number of documents assessed per query.

It should be noted that as **b** decreases, assessing documents becomes more productive (i.e. there are more relevant documents in the ranked list) and so assessing will begin to dominate the production process. That is, at some point the best strategy is to keep looking at more documents, rather than querying. This is because as **b** decreases the gain derived from assessing documents increases. Put more formally, as **b** tends to zero, then $\mathbf{A}^{1-\mathbf{b}}$ tends to **A**. For the above cost function, this point is when $\mathbf{b} \leq 0.5$, where from then on, the combination of inputs that minimizes the total cost is to issue one query, and then continue assessing until the desired level of gain is obtained. This can be seen as a boundary case. On the other hand, if **b** increases, then the combination of inputs that minimizes costs tends towards issuing many queries and assessing only one document per query (again this depends on the cost model). That is, if high precision queries are very cheap then it makes sense to keep issuing them until the desired level of gain is achieved. However, as the cost of a query increases there will be a point where it is better to substitute querying for more assessments per query. The **b** values shown in Table 1 show this trade-off. Note the **b** values shown are representative of standard IR algorithms [2].

This model of search behavior follows the basic economic principle that if cost goes up, consumption goes down [44]. It also reflects some of the empirical and simulated observations made in prior work [3, 7]. While this is promising, the model and its predictions are based on a number of assumptions which need to be considered.

Modeling Caveats and Assumptions Firstly, the model assumes that users are rational and will behave such that they would minimize interaction costs, and maximize performance. This is a common modeling assumption often employed (*c.f.* [4, 31, 32, 33, 34]). In the context of searching, an operation that users repeat and practice often, it has been shown that users adapt their behavior to systems [38], and that users do *try* to minimize effort and maximize performance [33] (i.e. they subscribe to Zipf's Principle of Least Effort). While the model may overestimate how well users could use an IR system, the assumption that they will try to optimize their behavior is, at least, reasonable. On an operational level, the model assumes that users will assess a fixed number of documents per query. However a user is likely to examine a different number of documents per query depending on the performance of the query. Given this assumption, it means that the model is rather coarse grained, considering the average number of documents assessed per query. A second operational point is that the cost model seems to ignore other costs, like viewing snippets, and how the costs of certain interactions may increase or decrease during the information seeking process. For example, a user may begin to run out of ideas for queries, and thus, increase the amount of time to generate subsequent queries. Ultimately, these simplifying assumptions reduce the problem to only the most salient factors, enabling us to reason about the relationship between the two main interactions within the search process: querying and assessing. The model generates a testable hypothesis, which we shall refer to as the *cost-interaction hypothesis*:

> As the relative cost of querying increases the average number of queries issued will decrease and the average number of documents assessed per query will increase.

4. METHOD

To test this hypothesis, we conducted an experiment, where we operationalized cost in terms of query time as done in [40]. Three interfaces were created that required different amounts of time to enter queries. Subjects were randomly assigned to use one of these interfaces (Figure 2):

1. Structured Query Interface (high cost);
2. Standard Query Interface (medium cost);
3. Query Suggestion Interface (low cost).

Aside from the different query facilities, these interfaces were similar and displayed 10 search results per page. Each query facility occupied the same amount of vertical space. To approximate the query cost in terms of time, we employed the GOMS Keystroke Level Model [8] using the timings from a search experiment by [39] (shown in Table 2). For this approximation, we assumed that the average length of a query was three terms, and each term was, on average, seven characters in length. A summary of the GOMS analysis for these query interfaces is provided in Table 3. This analysis is discussed in more detail in each sub-section below.

4.1 Standard Query Interface

The Standard interface functioned as the baseline and is similar to what is provided by modern search engines. We estimated that this interface would require a medium amount of query effort relative to the other interfaces. To issue a query using the Standard interface subjects need to: (1) Go

Figure 2: The Structured query interface (behind), the Standard query interface (middle) and the Query Suggestion interface (front).

to the search box, (2) Enter 3 query terms of seven characters in length (plus two spaces), (3) Submit the query by pressing the return key and then wait for a response. With respect to the GOMS analysis, the corresponding low-level actions for each of these steps are (1) MHPCH, (2) 3*7K+2K, (3) KR. Using the estimates shown in Table 2, the total amount of time is 10.9 seconds per query.

4.2 Structured Query Interface

The Structured interface consisted of 15 query boxes (3 rows x 5 columns), a search button and a plus button. Subjects could enter one term per box, but could not press the tab or enter keys to move among boxes. Each row of search boxes provided different Boolean functionality: AND, OR and NOT. While this interface might seem overly cumbersome, such search interfaces are common in commercial services; for example, EBSCO Host and ERIC provide a structured query interface similar to the one used in this study (see Hearst [22] for other examples). The plus button allowed query term boxes to be added. For this interface the GOMS analysis is as follows: (1) Go to a search box (MHPCH), (2) Type in the first query term (7K), (3) Go to the next search box (MHPCH), (4) type in the second query term (7K), (5) type in third query term (MHPCH + 7K),

Action	Time	Description
K	0.28	Press a key or button
P	1.1	Point with a mouse (excluding click)
H	0.4	Move hands to keyboard from mouse or vice versa
M	1.35	Mentally Prepare
R	0.8	System Response
C	0.2	Click

Table 2: GOMS Keystroke Level Model (time in seconds for each low level interaction)[8].

	Number of Queries Issued				
Interface	1	2	3	4	5
Structured	17.6	35.2	52.8	70.4	88.0
Standard	10.9	21.8	32.7	43.6	54.5
Suggestion	10.9	14.7	18.5	22.3	26.1

Table 3: Estimated GOMS total time spent querying in seconds for each interface.

and then (6) Submit the query - which requires the user to click the submit button (HC) and wait for response (R). The estimated total time required to enter each query is 17.6 seconds. To ensure that the physical costs associated with the Structured interface were as close to the GOMS model as possible, subjects had to retype query terms if they wanted to modify their previous query, i.e., query terms were removed from the boxes once the query was submitted, but the query was displayed on the screen. We assumed that this interface would also increase mental effort because subjects would have to understand how to enter terms.

4.3 Query Suggestion Interface

The Suggestion interface was identical to the Standard interface except query suggestions were presented after subjects entered their initial queries. For each topic, eight query suggestions were provided. These queries were collected in a previous study [26] and led to good results, where the queries retrieved between two and five TREC relevant documents in the top ten results. To issue a query using the Suggestion interface subjects needed to perform the same actions as for the Standard interface. However, for subsequent queries, subjects could select query suggestions, rather than type in queries. This resulted in the action sequence MHPCR taking approximately 3.8 seconds per suggestion. The Suggestion interface, while decreasing the amount of time to issue a query, should also reduce the mental effort associated with querying as it provides useful predefined suggestions. The query suggestions can be considered as a type of externalization, which is generally believed to reduce cognitive load in human-computer interactions [37]. Gwizdka's study of cognitive load in search found that an interface that displayed category terms reduced participants' cognitive load during query modification [20]. Thus, we anticipated that the query suggestion interface would also require less mental effort than the other two interfaces.

4.4 Corpus, Search Topics and System

A 3GB Text Retrieval Conference (TREC) test collection of over one million newspaper articles was used [45]. We selected three search topics from this collection: 344 (Abuses of E-mail); 347 (Wildlife Extinction) and 435 (Curbing Population Growth). We selected topics that had some contemporary relevance, that we thought would be of interest to

our target subjects and had a similar number of relevant documents available (123, 165 and 152, respectively). Our selection was also based on evidence from previous user studies with a similar system setup [26] where it was shown that the difficulty of these topics did not significantly differ. Subjects searched all three topics, where the topics were rotated with a Latin-square. The Whoosh IR Toolkit was used as the core of the retrieval system, with BM25 as the retrieval algorithm, using standard parameters, but with an implicit ANDing of query terms to restrict the set of retrieved documents to only those that contain all the query terms (similar to BM25A used in [2]). Subjects were not provided with a tutorial of the system.

4.5 Search Behaviors

To measure the impact of the cost variations on search behavior, the following signals were logged as subjects searched: start time of search tasks, end time of search tasks, queries issued, query suggestions used, results pages viewed, documents viewed, documents marked relevant, and task descriptions viewed. From the log it was possible to calculate the amount of time subjects spent issuing queries; examining search results pages and reformulating queries at this page; and viewing documents. It was also possible to examine features of the search interaction, such as, the number of terms per query, the depth of the last document viewed in the rank list, the number of queries issued and the number of documents saved.

4.6 Overall Workload: NASA TLX

The NASA Task Load Index (TLX) questionnaire was used to elicit subjects' perceptions of: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration. After finishing the three search tasks, subjects completed a NASA TLX questionnaire to rate their overall experience with the system, and then completed another NASA TLX questionnaire focused specifically on querying (Table 4). We reduced the number of scale points from 21 to 7 (the number of points on the computer version of the scale were difficult to distinguish and psychometric research shows that there is little gain in reliability beyond 11 points [30]) and we modified the factor statements so they matched the target task (in Hart's 2006 review of the usage of the NASA TLX it was noted that this modification was often performed by researchers [21]).

4.7 Subjects

Subjects were recruited from the University of Glasgow, UK[2]. Thirty-six undergraduate student subjects participated (12 subjects per interface). Twenty-one subjects were female and fifteen were male. Their average age was 21.8 (SD=4.4). Forty-seven percent were science majors and 53% were humanities majors. Subjects' search experience was measured using a modified version of the Search Self-Efficacy scale [13]. This instrument contains 14-items describing different search-related activities. Subjects respond to each item by indicating their confidence in completing each activity on a 10-point scale (1=totally unconfident; 10=totally confident). Subjects reported fairly high search self-efficacy [M=7.26 (SD=1.69)].

[2]Ethical Approval to conduct this study was obtained from the College of Science and Engineering, Reference No. CSE00913.

4.8 Instructions and Incentives

Subjects were instructed to imagine they were newspaper reporters and needed to gather documents to write stories about the provided topics. Subjects were told that there were over 100 relevant documents in the collection for each topic and they should try to find as many of these as possible during the allotted time (10 minutes per topic). To incentivize subjects, monetary bonuses were given to the top three performers for each topic per condition. Each subject was compensated with £10 and could earn an extra £2.50 per topic as a bonus.

5. RESULTS

Results are presented in three sections. The first two sections show how the search interface affected subjects' search behaviors, including the amount of time they spent engaged in different search processes. The third section presents results of the NASA TLX. We used ANOVA and post-hoc tests with Bonferroni's correction for analysis.

5.1 Search Behavior

The mean number of queries submitted by subjects, documents assessed per query, relevant documents found and assessment depth per query is shown in Table 5. Depth is the average position at which the last document was viewed in the search results list when at least one document was viewed.

	Interface		
Interactions	**Structured**	**Standard**	**Suggestion**
Queries (Q)	19.4 ± 10.6	35.0 ± 8.9	31.2 ± 10.4
Query Len.	3.7 ± 1.49	3.5 ± 0.7	3.2 ± 0.5
Assess. / Q	4.7 ± 4.2	1.6 ± 0.5	2.5 ± 1.5
Depth / Q	14.9 ± 12.2	5.0 ± 3.7	9.3 ± 10.1
Rels. Saved	47.7 ±24.1	34.7 ± 18.5	43.2 ± 18.7
TREC Rels.	17.3 ±8.1	11.5 ± 7.3	15.3 ± 6.7

Table 5: Search behavior and performance recorded per interface.

Subjects who used the Structured interface issued significantly fewer queries than those using the Standard and Suggestion interfaces ($F(2,33) = 7.59$, $p < 0.01$). There was no difference in query length across different interfaces. Subjects who used the Structured interface viewed significantly more documents per query and went to greater depths to view these documents. Both the number of documents examined per query ($F(2,33) = 3.26$, $p = 0.05$) and the depth per query were significantly different ($F(2,33) = 4.45$, $p < 0.01$), but only between the Structured and Standard interfaces. Subjects who used the Structured interface saved the most documents, followed by those who used the Suggestion interface and those who used the Standard interface, but these differences were not significant. Of these saved documents, the number relevant was determined by using the relevance judgments included in the TREC test collection. Subjects who used the Structured interface found more relevant documents than those who used the Suggestion and Standard interfaces. However, these differences were not significant.

5.2 Time Spent Engaged in Search Processes

Table 6 displays the average amount of time subjects spent issuing their first query, on search results pages (QSERPs)

Factors	Questions for: (a) System (b) Query
Mental Demand	How mentally demanding was it to (a) use this system to complete the search tasks?, (b) query?
Physical Demand	How physically demanding was it to: (a) use this system to complete the search tasks?, (b) query?
Temporal Demand	How hurried or rushed did you feel when (a) using this system to complete the search tasks?, (b) querying?
Effort	How hard did you have to work to (a) accomplish your level of performance with this system?, (b) query?
Performance	How successful were (a) you using this system to complete the search tasks?, (b) your queries?
Frustration Level	How insecure, discouraged, etc. were you while (a) using this system?, (b) submitting queries?

Table 4: Modified NASA TLX factor definitions for overall system load and query load.

and on document pages. The average amount of time spent on QSERPs includes the time spent viewing snippets and formulating queries. Subjects who used the Structured interface spent approximately 20 seconds longer formulating their first query than subjects who used the Standard and Suggestion interfaces ($F(2,33) = 4.05$, $p = 0.027$). Follow-up tests confirm that subjects who used the Structured interface took significantly more time to construct their initial queries than subjects who used the other two interfaces. These subjects also spent significantly more time on QSERPs, viewing snippets and reformulating queries ($F(2,33) = 8.149$, $p < 0.01$). There were no statistically significant differences in the amount of time subjects spent viewing documents among interface conditions, although subjects who used the Standard interface spent slightly more time per document.

	Interface		
Interactions	**Structured**	**Standard**	**Suggestion**
First Query	44.1 ± 35.7	22.9 ± 11.3	19.9 ± 12.3
QSERPs.	62.1 ± 32.9	28.6 ± 7.7	34.7 ± 16.1
Documents	15.1 ± 7.9	17.4 ± 5.1	15.3 ± 6.5

Table 6: Mean (SD) time (in seconds) spent formulating first queries, on search results pages viewing snippets and reformulating queries (QSERP) and viewing documents.

5.3 NASA TLX

Table 7 (top) displays the overall NASA TLX scores for each factor for each interface. These results capture the overall workload experienced by subjects as they engaged in all search behaviors (querying, viewing snippets and assessing documents). Subjects using the Standard interface reported experiencing the highest mental demand, followed by those using the Structured and Suggestion interfaces. This ordering was consistent for physical demand and temporal demands. Subjects indicated similar levels of success (performance), effort and frustration. None of the differences identified above were statistically significant.

Individual factor scores were summed to arrive at a total workload score. The Standard interface received the highest overall workload score, followed by the Structured and the Suggestion interfaces. These differences were also not significant. Table 7 also provides information about the relative contributions of each factor to overall workload for each interface. For example, effort was rated as the highest contributor to load by those who used the Structured and Suggestion interfaces, while mental demand was rated highest by those using the Standard interface. More interestingly, physical demand received the lowest scores regardless of interface. It was also the case that temporal demand was evaluated as the second highest contributor to load by subjects who used the Structured and Standard interfaces, but as the second lowest for those using the Suggestion interface.

Table 7 (bottom) displays the NASA TLX ratings for

System Load	Interface		
Factor	**Structured**	**Standard**	**Suggestion**
Mental	4.5 ± 1.4	5.1 ± 0.9	3.9 ± 1.6
Physical	2.3 ± 1.3	2.7 ± 1.1	1.9 ± 1.4
Temporal	4.6 ± 1.8	5.0 ± 1.4	3.6 ± 1.9
Performance	4.0 ± 1.4	4.1 ± 1.4	4.2 ± 1.4
Effort	4.8 ± 1.5	4.9 ± 1.4	4.7 ± 1.6
Frustration	3.5 ± 1.4	3.7 ± 1.7	3.8 ± 1.8
Total	23.7 ± 5.2	25.5 ± 5.9	22.1 ± 5.1

Query Load	Interface		
Factor	**Structured**	**Standard**	**Suggestion**
Mental	4.2 ± 1.4	4.8 ± 1.3	3.7 ± 1.9
Physical	2.2 ± 1.3	2.1 ± 1.2	2.2 ± 1.6
Temporal	4.2 ± 1.9	4.1 ± 1.8	3.0 ± 2.0
Performance	4.2 ± 2.0	3.1 ± 1.6	3.9 ± 2.0
Effort	4.4 ± 1.3	4.9 ± 1.6	4.5 ± 1.6
Frustration	4.0 ± 1.7	4.3 ± 1.5	3.1 ± 2.0
Total	23.2 ± 5.1	23.3 ± 7.0	20.4 ± 8.6

Table 7: Mean (SD) of NASA TLX Factors

query load. These results help us understand the load experienced by subjects as they queried. Subjects using the Standard interface reported experiencing the highest mental demand when querying, followed by those using the Structured and Suggestion interfaces. There were no differences among interfaces according to physical demand. With respect to temporal demand, subjects using the Structured and Standard interfaces reported higher demands than those using the Suggestion interface. Subjects who used the Structured interface described their queries as more successful (performance) than those using the Standard and Suggestion interfaces. Subjects using the Standard interface reported greater levels of frustration and effort. None of the differences identified above were statistically significant. With respect to overall query load, the Structured and Standard interfaces received similar scores, which were higher than overall query load for the Suggestion interface. These differences were not significant.

6. DISCUSSION

We found that subjects who used the Structured (high cost) interface submitted significantly fewer queries, examined more documents per query and went to greater depths in the search results list than subjects who used the lower cost Standard and Suggestion interfaces. This finding supports the *cost-interaction hypothesis* we generated from the microeconomic theory: as the cost of querying increases, the number of queries issued will decrease and the number of documents evaluated will increase. However, counter to what we expected, we found that subjects who used the Standard interface, which was constructed to represent medium cost, issued the most queries, viewed the fewest documents per query and were the shallowest in their evaluation of the

search results list. Our expectation was that these behaviors would be associated with the Suggestion interface since the cost of querying was considered lowest.

One possible explanation for this finding is that subjects who used the Suggestion interface may not have experienced any meaningful differences with respect to cost since they did not always click on the available suggestions. On average, subjects selected 7.31 query suggestions out of 24 query suggestions, which were available to them during their search sessions. The initial GOMS analysis assumed a greater uptake of suggestions, which would have resulted in much lower querying costs. If we revise our GOMS estimate, based on the actual usage statistics, then the time spent per query on the Suggestion interface would have been approximately, 9.24 seconds, on average[3]. This indicates that in terms of querying, the costs between these interfaces was very similar (and not as great as we originally anticipated). The empirical findings also confirmed that there was no significant difference in querying time between the Suggestion and Standard interface. However, the Suggestion interface may have introduced some added costs since subjects needed to examine and make decisions about the query suggestions. This is partially supported by the differences in the amount of time subjects spent on QSERPs evaluating snippets and formulating queries. Subjects who used the Suggestion interface spent slightly longer on QSERPs than subjects who used the Standard interface (34.7 and 28.6 seconds, respectively) and this increased time might have reflected time spent evaluating query suggestions. Though, here, the differences were not significant, which implies that the querying cost between these conditions was not different.

This raises a number of issues when applying the microeconomic theory in practice: (1) what costs are at play and how do we estimate them within the cost function, and (2) if the query costs were not significantly different, why do we observe different behavior between the Standard and Suggestion interfaces. Regarding costs, our research has shown that cost and how it is operationalized within the economic model needs to be reconsidered. Is the querying cost inclusive of the time spent viewing snippets or not, and if not, is it a document cost? If we use the QSERP estimates to denote the query cost (i.e. time spent querying and viewing snippets per query) then the Standard interface would be the least expensive: and our findings would be consistent with the cost-interaction hypothesis. However, the theory and the cost models employed would need to be refined, perhaps, redefined, to determine whether this is appropriate. We leave this to further work, and consider alternative explanations regarding the differences in search behavior between these two interfaces below.

We first considered whether the results with regard to the Standard and Suggestion interfaces were caused by differences in query quality and search result quality. Since the Suggestion interface provided good quality queries that had reasonably high precision, differences in query quality might explain why subjects who used the Suggestion interface tended to assess more documents per query and issue fewer queries than subjects using the Standard interface (i.e., they might have seen better search results). It is important to note that query quality differences would not have been caused by individual ability since subjects were randomly assigned to interface condition.

[3]Revised GOMS estimate: (23.89 queries entered manually * 10.9 seconds per query + 7.31 suggestions clicked * 3.8 seconds per suggestion) / (31.2 total queries) = 9.24 seconds per query issued.

	Structured	Standard	Suggestion
#Q	232	420	374
P@5	0.159	0.191	0.239**
P@10	0.139	0.171	0.217**
P@15	0.121	0.162*	0.196**

Table 8: Mean precision of queries issued across each interface. * (**) indicates significantly better than Structured (Structured and Standard).

To determine whether there was a difference in query quality, subjects' queries were submitted to the retrieval system and evaluated using TREC relevance judgements to better understand what types of performance subjects encountered. Table 8 reports the precision values at 5, 10 and 15 documents along with the number of queries issued in each group. The results show that the quality of the queries on the Suggestion interface was higher than the other two interfaces. Statistical testing revealed that system performance was significantly better across these three precision measures ($p < 0.05$). This finding suggests that subjects who used the Suggestion interface issued better quality queries and potentially encountered more TREC relevant documents in the ranked lists, but we note that there were no significant differences in the number of documents subjects saved or the number of TREC-relevant saved, so we cannot conclude that subjects using the Suggestion interface experienced better performance, especially considering the work of Turpin and Hersh [43] who have shown that system performance evaluated in this way does not always map to user performance.

The results from the NASA TLX also provide additional insights about how subjects experienced these interfaces, and the extent to which cost, as we have manipulated it in this study, impacted their experiences. Although many interesting differences are discussed below, it is important to keep in mind that none of these were statistically significant, so this discussion is primarily presented in the service of future research.

The NASA TLX was administered twice: once to focus on system load and once to focus on query load. The system load allowed us to understand the entire experience (querying, evaluating snippets and results, search result quality), while query load isolated the load introduced by the query facilities. One of the most interesting and consistent findings was that physical demand contributed the least to subjects' overall loads. For query load, the Structured, Standard and Suggestion interfaces were rated nearly the same for this factor (2.2, 2.2, and 2.1, respectively). There were greater differences in subjects' ratings of physical demand for system load, with the Standard interface receiving a rating of 2.7, followed by Structured (2.3) and Suggestion (1.9) interfaces. These ratings were, in general, low, when compared to the other factors, which suggests that our manipulation of physical demand may not have been as extreme as expected.

Results also showed that subjects associated the most mental demand with the Standard interface (4.8), followed by the Structured (4.2) and Suggestion (3.7). This ordering was consistent for the mental demand associated with system load, although the values were slightly larger (5.1, 4.5 and 3.9, respectively). Compared to the ratings for physical demand, it is clear that mental demand contributed more

to subjects' overall workloads. The Suggestion interface received the lowest ratings for mental demand as expected, but the Standard received higher ratings than the Structured, which was unexpected. We believe that the higher mental demand associated with the Standard interface was a result of subjects creating more queries while searching.

While one might argue that the Structured interface is clunky and less usable than the other interfaces, subjects did not report high levels of frustration with the query facility. Rather, the highest levels of frustration were reported by subjects who used the Standard interface. This suggests that subjects did not find the added time costs associated with the Structured interface annoying and that subjects were more frustrated by having to enter more queries with the Standard interface. The Suggestion interface, as expected, received the lowest frustration rating for query load. These results make us question how subjects would have rated the usability or aesthetics of the interfaces. We hypothesize that subjects would rate the Structured interface the lowest, which is interesting since they actually performed better with it. We leave this for future work.

A final difference we observed in the NASA TLX was with respect to the temporal factor, which gauged how hurried or rushed subjects felt when doing the search tasks. Subjects who used the Standard interface reported experiencing the highest levels of temporal demand (5.0), followed by those who used the Structured interface (4.6) and those who used the Suggestion interface (3.6). This difference suggests that a system that encourages the type of search behavior that subjects who used the Standard interface engaged in - issuing more queries, evaluating fewer documents per query and shallowly evaluating search results lists - might have negative psychological consequences, assuming that feeling hurried and rushed generates stress. There were no differences between the Standard and Structured interfaces for temporal demand for query load, which seems to further suggest that it was the total search strategy engaged in by those who used the Standard interface that contributed to the differences in overall temporal demand.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated how query cost affects search behavior. We used microeconomic theory to motivate our study and conducted a theoretical analysis, which generated the *cost-interaction hypothesis*. A laboratory study with 36 subjects was conducted to evaluate this hypothesis using three interfaces: Structured, Standard and Suggestion. We found partial support for this hypothesis: subjects who used the high cost Structured interface submitted fewer queries, spent more time on search results pages, examined more documents per query, and went to greater depths in the search results list than subjects who used the lower cost Standard and Suggestion interfaces.

Our results have both theoretical and practical implications. Attempts to formally model search interaction are promising, but the results of this study suggest that refinements to the microeconomic theory are required to improve the realism of the model. Our results also imply that at least one additional factor should be included in the gain and cost functions to account for viewing snippets, and that more sophisticated cost functions may be more appropriate. We found that in testing the theory, care needs to be taken to control each of the factors involved. Changing the costs may change other aspects of the interaction and these need to be accounted for when testing hypotheses generated by the model.

Future work will focus on further refining the search economic theory/models and exploring how alternative designs for query facilities and other search interface features might encourage users to engage in and adopt more positive and successful search behaviors.

Acknowledgments We would like to thank Kelly Marshall for her help in conducting the user experiments.

8. REFERENCES

[1] E. Agapie, G. Golovchinsky, and P. Qvardordt. Encouraging behavior: A foray into persuasive computing. In *Proceedings on the Symposium on Human-Computer Information Retrieval*, 2012.

[2] L. Azzopardi. The economics in interactive information retrieval. In *Proceedings of the 34th ACM conference on research and development in information retrieval (SIGIR)*, pages 15–24, 2011.

[3] 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.

[4] M. J. Bates. Training and education for online. chapter Information search tactics, pages 96–105. Taylor Graham Publishing, London, UK, UK, 1989.

[5] N. J. Belkin, D. Kelly, G. Kim, J.-Y. Kim, H.-J. Lee, G. Muresan, M.-C. Tang, X.-J. Yuan, and C. Cool. Query length in interactive information retrieval. In *Proceedings of the 26th ACM conference on research and development in information retrieval (SIGIR)*, pages 205–212, 2003.

[6] N. J. Belkin and A. Vickery. *Interaction in Information Systems*. University Press, 1985.

[7] J. Brutlag. Speed matters for google web search. In *Technical Report, Retrieved online at http://googleresearch.blogspot.com/2009/06/speed-matters.html, on May 11, 2013*, 2009.

[8] S. K. Card, T. P. Moran, and A. Newell. The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7):396–410, 1980.

[9] M. D. Cooper. A cost model for evaluating information retrieval systems. *Journal of the American Society for Information Science*, pages 306–312, 1972.

[10] W. S. Cooper. On selecting a measure of retrieval effectiveness, part 1: The subjective philosophy of evaluation. *Journal of the American Society for Information Science*, 24:87–100, 1973.

[11] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practice*. 2009.

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

[13] S. Debowski, R. Wood, and A. Bandura. The impact of guided exploration and enactive exploration on self-regulatory mechanisms and information

acquisition through electronic enquiry. *Journal of Applied Psychology*, 86:1129–1141, 2001.

[14] 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.

[15] 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, Jan. 2002.

[16] G. Dupret and B. Piwowarski. A user behavior model for average precision and its generalization to graded judgments. In *Proceedings of the 33rd ACM conference on research and development in information retrieval (SIGIR)*, pages 531–538, 2010.

[17] K. Franzen and J. Kalgren. Verbosity and interface design. In *SICS Technical Report: T2000:04, Retrieved online at http://soda.swedish-ict.se/2623/2/irinterface.pdf on May 11, 2013*, 1997.

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

[19] 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.

[20] J. Gwizdka. Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology*, 61(11):2167–2187, Nov. 2010.

[21] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, pages 904–908, 2006.

[22] M. A. Hearst. *Search User Interfaces*. New York, NY: Cambridge University Press, 2009.

[23] 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.

[24] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of information retrieval technology. *ACM Trans. on Information Systems*, 20(4):422–446, 2002.

[25] D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3:1–224, 2009.

[26] D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd ACM conference on research and development in information retrieval (SIGIR)*, pages 371–378, 2009.

[27] 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.

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

[29] 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.

[30] J. C. Nunnally. *Psychometic Theory*. McGraw, 1978.

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

[32] P. Pirolli and W. T. Fu. Snif-act: a model of information foraging on the www. In *Proceedings of the 9th International Conference on User Modeling*, pages 45–54, 2003.

[33] P. Pirolli, P. Schank, M. Hearst, and C. Diehl. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the ACM SIGCHI Conference*, pages 213–220, 1996.

[34] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *Proceedings of the INTERACT and SIGCHI Conference*, pages 269–276, 1993.

[35] I. Ruthven. Interactive information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 42:43–92, 2008.

[36] G. Salton. Evaluation problems in interactive information retrieval. *Information Storage and Retrieval*, 6:29–44, 1970.

[37] M. Scaife and Y. Rogers. External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, 45:185–213, 1996.

[38] C. L. Smith and P. B. Kantor. User adaptation: good results from poor systems. In *Proceedings of the 31st ACM conference on research and development in information retrieval (SIGIR)*, pages 147–154, 2008.

[39] M. D. Smucker. Towards timed predictions of human performance for interactive information retrieval evaluation. In *Proceedings of the Symposium on Human-Computer Information Retrieval*, 2009.

[40] 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.

[41] L. T. Su. Evaluation measures for interactive information retrieval. *Information Processing and Management*, 28:503–316, 1992.

[42] 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.

[43] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th ACM conference on research and development in information retrieval (SIGIR)*, pages 225–231, 2001.

[44] H. R. Varian. *Intermediate microeconomics: A modern approach*. W.W. Norton, New York:, 1987.

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

[46] E. Yilmaz, M. Shokouhi, N. Craswell, and S. Robertson. Expected browsing utility for web search evaluation. In *Proceedings of the 19th ACM conference on Information and knowledge management (CIKM)*, pages 1561–1564, 2010.