Using Information Scent and Need for Cognition to Understand Online Search Behavior

Wan-Ching Wu
SILS
University of North Carolina
Chapel Hill, NC, 27599 USA
wanchinw@email.unc.edu

Diane Kelly
SILS
University of North Carolina
Chapel Hill, NC, 27599 USA
dianek@email.unc.edu

Avneesh Sud
Microsoft
One Microsoft Way,
Redmond, WA, 98052 USA
avneesh.sud@microsoft.com

ABSTRACT
The purpose of this study is to investigate the extent to which two theories, *Information Scent* and *Need for Cognition*, explain people's search behaviors when interacting with search engine results pages (SERPs). Information Scent, the perception of the value of information sources, was manipulated by varying the number and distribution of relevant results on the first SERP. Need for Cognition (NFC), a personality trait that measures the extent to which a person enjoys cognitively effortful activities, was measured by a standardized scale. A laboratory experiment was conducted with forty-eight participants, who completed six open-ended search tasks. Results showed that while interacting with SERPs containing more relevant documents, participants examined more documents and clicked deeper in the search result list. When interacting with SERPs that contained the same number of relevant results distributed across different ranks, participants were more likely to abandon their queries when relevant documents appeared later on the SERP. With respect to NFC, participants with higher NFC paginated less frequently and paid less attention to results at lower ranks than those with lower NFC. The interaction between NFC and the number of relevant results on the SERP affected the time spent on searching and a participant's likelihood to reformulate, paginate and stop. Our findings suggest evaluating system effectiveness based on the first page of results, even for tasks that require the user to view multiple documents, and varying interface features based on NFC.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - search process.

Keywords
Search Stopping; Search Behavior; personality; Need for Cognition; Information Scent; Search Depth; Query Reformulation; Pagination; Search Strategies.

1. INTRODUCTION
Online searching can be conceptualized as an iterative decision-making process. After an initial query is submitted, a searcher examines a search engine results page (SERP) and decides to select one or more of the results, paginate to the next SERP, issue another query or end their search task. While many studies have used search log analysis to model these types of search behaviors (e.g., [16] [42]), the goal of many of these studies has been to learn from large scale data a parsimonious set of the most important features to predict future search behavior. While these studies provide much in the way of *description* and *prediction*, they offer few *theoretical explanations* as to *why* people engage in certain search behaviors. In this study, we investigate the extent to which two theories, *Information Scent* [33] and N*eed for Cognition* [6], explain people's search behaviors when interacting with SERPs.

Information scent is part of *Information Foraging Theory,* which compares information-seeking behavior to the food foraging behaviors of animals [33]. Just as animals use environmental cues to identify the most useful places to forage for food, human beings look for information from sources they believe are the most cost-effective by making predictive judgments using proximal cues. Examples of proximal cues include snippets, thumbnails and URLs. These proximal cues provide *information scent,* which emphasizes their utility in suggesting the relevance of some distal content. While previous work has examined the relationship between the information scent of individual search result surrogates (i.e., snippets) on search behavior [25][29], in this work, we propose that the *initial* search result *page* can be viewed as a surrogate for the *entire set of results* retrieved for a query, and examine how the *information scent level,* or the number of relevant results*,* and *information scent pattern,* or the distribution of relevant results*,* of the initial SERP influence people's search behaviors.

In this study, we also investigate how *Need for Cognition* (NFC), a personality trait that measures the extent to which a person enjoys cognitively effortful activities, impacts search behaviors. Interest in personality-based designs has been growing in the HCI community and is motivated by the assumption that people with different personality traits will respond differently to design cues and interact in different ways with systems [32]. While many studies in information science have investigated the effect of individual differences such as cognitive styles, gender and age on search behavior (e.g., [17]), the relationship between personality and search behavior has received relatively less attention, especially in IR. Moreover, several IR evaluation measures include parameters that can be tuned based on individual user characteristics such as patience (e.g., nDCG [23]). However, no research has been done to investigate potential characteristics that might impact a person's willingness to examine a search results list. One possible personality trait from the psychological literature that may influence search behavior is *Need for Cognition* (NFC). Research has shown that people with higher NFC spend more time evaluating ideas and information, and solving problems [8]. Given that search is a cognitive activity and that people with higher levels of NFC spend more time processing and evaluating information, we were interested in examining whether and how differences in NFC

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07…$15.00.
http://dx.doi.org/10.1145/2600428.2609626

impact how many queries a person enters, how deep they go in the search results list and when they decide to stop searching.

2. RELATED WORK

The section reviews work related to the search behaviors we are most interested in: query reformulations, search depth, and search stopping. This is followed by an overview of research related to information scent and need for cognition.

2.1 Query Reformulation

Literature on query reformulations can be divided into computer-generated query suggestions and query rewriting, and human-generated reformulations. This section is dedicated to the latter, with a focus on Web search reformulation behavior. There have been a number of studies that have analyzed queries and query reformulations, and created classifications of query reformulation strategies. For example, Anick [1] classified 100 query reformulations into 11 functional categories, such as location, acronym, and alternative. Rieh and Xie [36] analyzed 313 Web search logs and reported three facets of query reformulations: content, format and source. Huang and Efthimiadis [19] proposed a taxonomy based on the lexical qualities of queries and evaluated the effectiveness of different query reformulation strategies. They concluded the effectiveness of query reformulation strategies depends on the user's reaction to the first set of results retrieved.

Spink [39] analyzed Excite search logs and found that users typically entered 2.84 queries per session, and in about two-thirds of the sessions, there were query reformulations. Jansen and Spink [22] surveyed nine search engine logs from 1997 to 2002 and found that there was an increase in query reformulations. Jansen et al. [21] analyzed search logs collected from 2005 and found that about 40% of query occurrences were reformulations, yet it was unclear how these reformulations were connected to user sessions. Recently, Hassan et al. [18] used clicks and query reformulations as indicators of search satisfaction. The researchers found query reformulations were a strong indicator of task difficulty and task failure. Moreover, queries associated with unsuccessful tasks were more similar to one another than queries associated with successful tasks. These findings provide useful perspectives for understanding the relationships between query reformulations and stopping.

2.2 Search Depth

Studies about search depth have reported the depth on a ranked search result list at which searchers stop clicking, mouse hovering, or gazing before re-querying. A series of studies from the early 2000's examined queries from the Excite search engine and observed that people typically examined 1.7 pages per query and for about 50% of the queries, searchers went to the next page before reformulating their queries [39]. The researchers also found that search depth decreased over time; the trend of viewing only the initial SERP increased from 29% to 73% in US-based search engines from 1997 to 2002 [22]. Pagination has rarely been observed in recent studies. Cutrell and Guan's eye tracking study reported that people examined the first eight results before they re-issued another query [14]. Lorigo et al. [28] examined subjects' scan paths as they carried out search tasks. They found that on average subjects scanned just 3.2 distinct search results for each query. Using cursor movements, Huang, White, and Dumais [20] found that people re-queried after inspecting the top four results. Another experimental study motivated by Search Economy Theory found that search depth was affected by query cost [2]. Subjects that used an interface, which required more time to enter a query, entered significantly fewer queries and went to greater depths in the search results list than subjects who used a standard search interface. These results suggest that certain aspects of the search interface can impact search behavior and also provide a theoretical explanation for this behavior.

2.3 Search Stopping

Search stopping is generally used to describe the point a person decides he or she has enough information to complete his or her search task. Terms such as stopping behavior, search termination, searcher patience or search persistence have been used to describe this phenomenon, but no formal definition or models have been provided in the literature. To date, the search stopping literature covers issues related to stopping rules for determining when people have found enough information or when they declare search a failure [5][12][13][28][31][34], search patterns and effort preceding a stopping point [30][40], the relationship between situational and individual differences on stopping [34][44], and the consequences of different search stopping decisions [29]. The majority of these studies have used interviews to understand when people decide to stop and have concluded that stopping decisions are based on intuition, or "the feeling of good enough" [44]. Only a few studies have attempted to quantitatively characterize and predict stopping. For example, Kraft and Lee [27] modeled expected search length through mathematical approximations using stopping rules including the satiation rule, the disgust rule, and the combination rule. This study, though, was in the context of batch-mode searching, rather than interactive searching. More recently, Toms and Freund [40] studied actions that preceded stopping in online information seeking in order to predict stopping and found that the most prevalent stopping pattern included issuing a query, examining results and viewing a page. Dostert and Kelly [15] found that when participants stopped they believed they found most of the relevant information, but in reality they had only identified 51-60% of relevant information. This study was one of the first attempts to quantify the sense of "feeling good enough."

2.4 Information Scent

Information scent, a component of Information Foraging Theory, is the subjective perception of the value and cost of information sources from proximal cues, such as search result snippets representing the page content [10][33]. Pirolli and his colleagues have used information scent in a number of studies to better understand web search behavior [9][10][11]. For example, Chi et al. [10][11] developed and evaluated a computational method to predict surfing paths based on given information needs and starter pages. Card et al. [9] observed that if a person started with a high information scent web page, he or she would visit more web pages at the site. They also found that as the information scent of web pages declined, there was a tendency for the person to leave the site or return to a previously visited page. Information scent has also been used to better understand how people evaluate search results. Loumakis et al. [29] investigated how the information scent associated with images on SERPs impacted evaluation behavior. They found that when images were added to text snippets, regardless of image quality, participants were more confident they could find an answer. However, the researchers studied simple fact-finding tasks and did not investigate the effects of information scent on search behavior. Kammerer et al. [25] found that by adding source cues to search result snippets, searchers paid less attention to commercial search results and selected more results from authoritative sources than when source cues were not available. However, participants were provided 30 results to "select" from rather than to "search;" therefore, the effect of information scent on search behaviors in relation to SERPs remains unclear.

2.5 Need for Cognition (NFC)

Considerable research related to NFC has been conducted in the fields of social psychology, personality psychology, behavioral medicine, education, media studies, marketing and law. In general, this research has found that high NFC is associated with higher motivation to seek information, increased information processing effort, and an increased ability to assess message quality. Bailey [3] found in a study during which managers were asked to evaluate job candidates, that high NFC managers evaluated candidates' information more thoroughly than low NFC managers. In another study where students were given editorials to evaluate, high NFC students performed better at discriminating between strong and weak arguments than low NFC students [8]. This evidence suggests NFC may also explain variations in the amount of information searched for and the strategies people use when conducting information search. To date, the only study that has investigated the role of NFC in IR was Scholer, et al. [37] who investigated if this characteristic mediated the extent to which a list of documents, with varying densities of relevant and non-relevant documents, impacted people's relevance judgments of those documents. While the researchers did not find that NFC mediated this relationship, their study participants did not differ greatly with respect to NFC, so lack of variance on this measure might have prevented them from observing an impact.

3. RESEARCH QUESTIONS

This study is concerned with using information scent and need for cognition to explain the search behavior of people conducting open-ended search tasks, or tasks that require several documents to complete, and addresses the following three research questions:

RQ1: What is the relationship between the information scent level of the first SERP and search behaviors?

Searchers often have to issue several queries before they obtain a sufficient amount of information for open-ended search tasks. It is proposed that the first SERP can be viewed as a surrogate for the entire set of results returned in response to a query. Just as searchers can be made aware of the potential value of a single search result by the information scent of the snippet [25][29] and the potential usefulness of an entire website based on its homepage [9], arguably they may also attempt to predict the potential value of the entire set of results retrieved for a query based on the quality of the initial SERP. Based on the same analogy, the information scent of the first SERP can possibly be used to predict how long a searcher will evaluate a set of search results for a single query. If the *number* of relevant results is higher on the first page, this might increase the interactions with the result set compared to when the first SERP has fewer relevant results.

RQ2: What is the relationship between the information scent pattern of the first SERP and search behaviors?

The finding that when information scent declined on a web page, searchers tended to leave the website [9] also suggests that the *distribution* of relevant results might impact how long a searcher interacts with a set of results. It is proposed that when relevant search results are evenly distributed across the first SERP, a certain degree of information scent is maintained throughout the first SERP. This in turn, might induce a searcher to interact with the result set to a greater extent. In contrast, when relevant results are concentrated in batches on the first SERP, searchers might be more likely to believe they have seen all the relevant results.

RQ3: What is the relationship between NFC and search behaviors?

Individuals with high NFC have been found to exert more effort during information processing. This has two possible implications for information search. First, people with high NFC may examine more information for a given query. Since high NFC searchers enjoy the thinking process, their higher motivation to process information may allow them to be more resilient to non-relevant search results, thus lowering their motivation to reformulate. Second, it is also likely that people with high NFC may exert more effort on query reformulation. Since people with high NFC enjoy thinking and query reformulation is a cognitive task, effort may manifest in more frequent query reformulations rather than prolonged engagement with search results. Support for the latter hypothesis also comes from research that has demonstrated that high NFC people make more accurate judgments about message quality [8]. If high NFC people are more capable of discriminating high quality from low quality content, they may reformulate as soon as they encounter a bad document in search for higher quality information.

4. METHOD

A laboratory experiment was conducted with three independent variables: information scent level (ISL), information scent pattern (ISP) and need for cognition (NFC). The first two variables were within-subject variables, while the last was a between-subjects variable. Information scent level (ISL) was defined as the number of relevant documents on the first SERP and was operationalized with three levels: high, medium and low (Table 1). In the high condition, five relevant documents appeared on the first SERP, in the medium, three relevant documents and in the low, one relevant document. To control for the influence of result positioning, relevant results were always placed consecutively starting from the first result and also represented the best possible orderings for each level. Information scent pattern (ISP) was defined as the distribution of relevant documents on the first SERP and was operationalized with three levels: persistent, disrupted and bursting, each of which differed according to the distribution of four relevant search results on the first SERP (Table 1). The persistent pattern initially has a high scent, with relevant documents in the first two positions, followed by two more relevant documents at a consistent interval. The disrupted pattern starts with a strong scent, which quickly extinguishes. This represents the best possible ordering for four relevant documents. The bursting pattern represents a case where the scent is not initially present, but then appears strong and steady in the middle of the list before extinguishing.

Table 1. Information Scent Level and Information Scent Pattern conditions (R=Relevant; NR=Non-Relevant)

	Level			Pattern		
Rank	Low	Medium	High	Persistent	Disrupted	Bursting
1	**R**	**R**	**R**	**R**	**R**	NR
2	NR	**R**	**R**	**R**	**R**	NR
3	NR	**R**	**R**	NR	**R**	NR
4	NR	NR	**R**	NR	**R**	**R**
5	NR	NR	**R**	**R**	NR	**R**
6	NR	NR	NR	NR	NR	**R**
7	NR	NR	NR	NR	NR	**R**
8	NR	NR	NR	**R**	NR	NR
9	NR	NR	NR	NR	NR	NR
10	NR	NR	NR	NR	NR	NR

During the experiment, a total of six search tasks were assigned to each participant to complete. The study tasks were rotated according to a Latin Square design. In three of the six tasks, ISL was manipulated while in the other three tasks ISP was manipulated. Participants were made to believe they were using an experimental search engine and were asked to enter self-generated queries to complete the tasks. However, no matter what queries they issued for their first three query submissions for a given task, they received preselected search results that reflected various ISL and ISP conditions. The word "search result set" is used to describe the entire preselected search results, which contained a total of ten search result pages (100 results). The manipulation is summarized in Figure 1. From left to right three preselected search results sets were presented to participants for each query they submitted. In each search result set, the very first SERP was manipulated to reflect a specific experimental condition. In the three tasks where ISL was manipulated, the first three SERPs displayed high, medium and low ISL conditions. In the other three tasks where ISP was manipulated, the three first SERPs displayed persistent, disrupted and bursting ISP conditions. The orders of these conditions were rotated for each task, so that each participant experienced each condition on the first SERP for one task. Results presented at the 11th to 100th positions for the first three sets of results were also preselected. Only the twelfth, fifteenth and eighteenth results on the second SERPs were relevant; we did not include more relevant documents on this page because we were interested in focusing on participants' transitions from the first SERP. However, we included some relevant documents because we did not want participants who went to these subsequent pages to learn from their interactions that paginating to the second page always ended up futile. They were not required to view all SERPs or enter any pre-specified number of queries. If a participant submitted more than three queries, the Bing search API [1] was used to fetch results.

Need for Cognition (NFC), the other independent variable, was measured by the NFC scale [7]. The NFC scale contained 18 items assessed with a 5-point scale (5=extremely characteristic of me; 1=extremely uncharacteristic of me). NFC scores for participants were derived by averaging responses to these 18 items. Participants completed this scale at the end of the study.

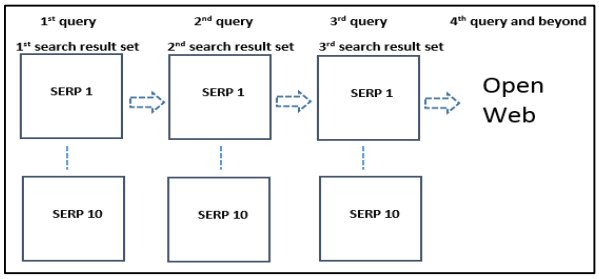


Figure 1. Search result evaluation flow

4.1 Tasks & Documents

We reused and adapted a set of tasks that had been used in a previous project so that we had more evidence about the range and types of queries people might submit [43]. We did not want to assign search tasks where people would enter a lot of different queries as we wanted to ensure that our preselected results would be relevant. The query logs from the previous study showed that most subjects submitted keywords from task descriptions and thus we believed potential queries were within predictable ranges. This also allowed us to use some of the documents retrieved by previous participants in this study. The tasks were open-ended search tasks that presented a motivation for seeking the information and several questions of interest. All tasks took this basic form and asked participants to gather and analyze information, and make a decision. An example task can be seen in Figure 2; all tasks are available online[2]. Participants were told to spend as much as they wanted on each task, until they collected a set of pages that in they felt satisfied the information requirement of the task. Participants were not required to produce any answers, but were told to save pages that helped them address the task.

Special care was taken to ensure that participants would experience the preselected search results in accordance to the manipulated information scent levels and patterns. Relevant search results were selected from the clicked webpages gathered by participants in the previous project. Each clicked page was reviewed by one of the study authors and two other assessors. Only webpages that were judged by all three assessors as relevant were used. Non-relevant search results were identified by submitting queries composed of a keyword from the task description and some terms unrelated to the task to a popular search engine. For example, non-relevant search results for one task which was about methods of tattoo removal were gathered by submitting the queries *tattoo designs* and *tattoo mistakes*. It was also important to ensure that the result snippets clearly reflected whether a landing page was relevant or not so that subjects would experience the intended information scents. Non-relevant search results and snippets were also evaluated by two assessors. Eight pilot studies were conducted to ensure that all the manipulations worked and did not cause suspicion.

4.2 Search Interface

The search interface is shown in Figure 2. At the start of the task, the task description was shown at the top along with a query box. After participants submitted their initial queries, a page of ten results was displayed. When a participant clicked on a search result, the landing page was presented in a separate tab and participants were asked whether they wanted to save the page (Figure 3). Once participants submitted a response, the tab automatically closed and participants were taken back to the SERP. If participants attempted to close the tab without answering the question, a warning message appeared. Participants clicked "Done" in the upper right corner of the search page when they finished the task. Before starting the study, participants were given a practice task.

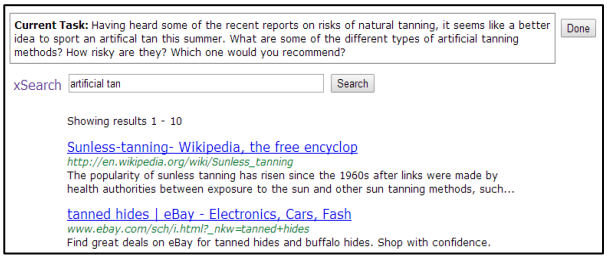


Figure 2. Search Interface

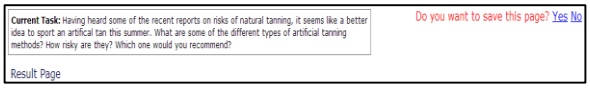


Figure 3. Landing Page Interface

[1] http://datamarket.azure.com/dataset/bing/search

[2] http://wanchinw.web.unc.edu/sigir-supplementary-page/

4.3 Search Behaviors

The dependent measures consisted of search behaviors, which were recorded in a search log:

- **QueryAction**: a categorical measure of the outcome after a query submission. The measure has three values: query reformulation on the first SERP, paginating to the second SERP or stopping a task on the first SERP.
- **Abandonment:** not clicking on a SERP after a query submission.
- **NumPagination:** frequency of pagination within a search result set.
- **NumQuery**: number of query submissions during a task.
- **Time**: amount of time spent examining a search result set.
- **DeepestRankClick**: deepest rank of a clicked result.
- **DeepestRankHover:** deepest rank where a mouse hover was observed.
- **NumExamined**: number of documents examined for each search result set.
- **NumPred, NumRele and NumNonRele:** A *predictive* judgment of relevance is made when a searcher views only a search result snippet, while an *evaluative* judgment of relevance is made after a searcher examines the content [35]. NumPred was the number of documents participants clicked on but decided not to save; NumRele was the number of documents that were clicked and saved; NumNonRele refers to the number of documents that were ranked before DeepestRankClick but not clicked.

Participants' searches were also captured with Morae and at the end of the study they were interviewed using stimulated recall with video recordings of three of their searches.

4.4 Participants

Participants were recruited by sending an email to the staff mailing list at our university. Forty-eight people participated, but only data from 47 participants are included because of a logging failure. Participants' average age was 38.29 (range: 19-65). Their job titles included web developer, HR specialist, financial aid counselor, administrative assistant, librarian, lab manager, instructor, research assistant, play writer, fire department technician, and sales manager. Participants scored an average of 7.81 (*SD*=1.34) on a 10-point search self-efficacy scale, showing a medium to high level of confidence in web search skills. Participants were paid $20 cash for their participation.

4.5 Manipulation Check

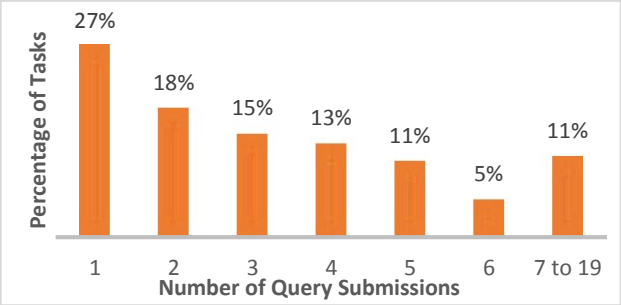
During each search task, participants were shown preselected search result sets for their first three queries. Among the documents participants clicked on for their first three queries, 98.22% were judged relevant by the assessors and among those they saved, 99.22% were judged relevant by the assessors, which shows the manipulation of document relevance was successful. To examine whether our pre-selected search results caused any suspicion, we asked participants at the end of the experiment to comment on the quality of the search results. Most participants reported the quality was good. Some commented there were many non-relevant results on the first page, but explained this by the popularity of certain webpages or advertisements or attributed this to their own ambiguous queries. No participant indicated they suspected manipulation.

5. RESULTS

5.1 Overview

The forty-seven participants completed a total of 282 tasks. Participants were able to enter as many queries as desired. Figure 4 shows the distribution of the number of queries submitted per task. In about 60% of the tasks, participants submitted 1-3 queries.

Figure 4. Query submission frequency per task



Descriptive statistics for continuous search behavior measures aggregated at the task level can be seen in Table 2. "*n*" next to the mode represents the frequency of tasks for the most common action. On average, participants issued 3.47 queries per task, and for 76 tasks, participants issued only one query. Forty-six out of the 76 tasks were ISP tasks (four relevant results were presented on the first SERPs), and in another 19 tasks participants encountered a high ISL (five relevant results on the first SERPs); moreover, 32 out of the 76 tasks involved at least one pagination. This suggests that encountering more relevant documents during a query submission probably resulted in satisfaction without reformulation even when the assigned tasks were open-ended tasks. Participants also paginated an average of 1.47 times per task; however, in more than half of the tasks they never paginated, and as many as twelve participants never paginated during the entire experiment. Tasks lasted 6.6 minutes on average with a large range: the minimum time was 76 seconds and the maximum, 19 minutes. Participants saved 4.82 results per task.

With respect to NFC, our participants scored an average of 3.75 on a five-point scale (*SD*=0.55) and the distribution of NFC scores was normal. We correlated NFC with search behavior measures aggregated at the task level to examine whether there were any relationships and found that participants with higher NFC paginated less than those with lower NFC (r= -.33, p=.023, N=47).

During the three search tasks where ISL was manipulated, participants were exposed to 105 low, 109 medium, and 126 high result manipulations, while in the other three tasks where ISP was manipulated participants were exposed to 98 persistent, 101 disrupted, 104 bursting result manipulations. The outcomes of

Table 2. Search behaviors at the task level

	Time (sec)	NumQuery	NumPagination	NumExamined	DeepestRankClick	DeepestRankHover	NumPred	NumRele	NumNonRele
Mean	332	3.47	1.47	6.7	17.28	33.22	1.76	4.82	10.59
Median	291	3	0	6	14	19	1	4	8
Mode	--	1 (*n*=76)	0 (*n*=167)	4 (*n*=56)	12 (*n*=17)	19 (*n*=24)	0 (*n*=98)	4 (*n*=71)	0, 1, 5 (*n*=66)
SD	427	2.67	2.97	3.96	14.36	35.15	2.28	2.70	11.79

participants' query submissions are presented in Figure 5. From low to high ISL, reformulations decreased and stoppings increased. When comparing ISP, bursting appeared to lead to the highest percentage of reformulations and the lowest percentage of stoppings. The distribution of behaviors for persistent and disrupted were similar. Pagination remained relatively constant across all ISL and ISP conditions.

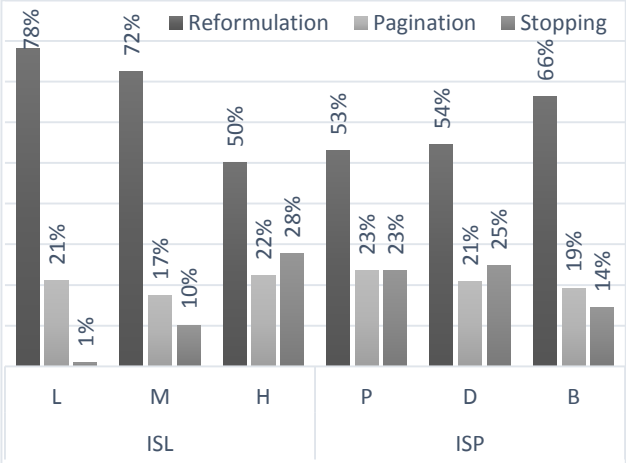


Figure 5. Reformulation, pagination and stopping by experimental condition (L=low; M=medium; H=high; P=persistent; D=disrupted; B=bursting).

We also examined query abandonment to understand how ISL and ISP affected participants' reactions to result snippets. The bars in Figure 6 represent the abandonment rate given each condition. When participants were presented with a SERP with low ISL, they chose to leave without examining any document around 42% of the time, while the abandonment rate for high ISL was only 1.6%. The differences among ISP conditions were not as dramatic but abandonment for the bursting condition happened 10% more than in the persistent condition, which is interesting since these condition had the exact same number of relevant documents. Also interestingly, this abandonment rate was higher than that of medium ISP, which had one less relevant result.

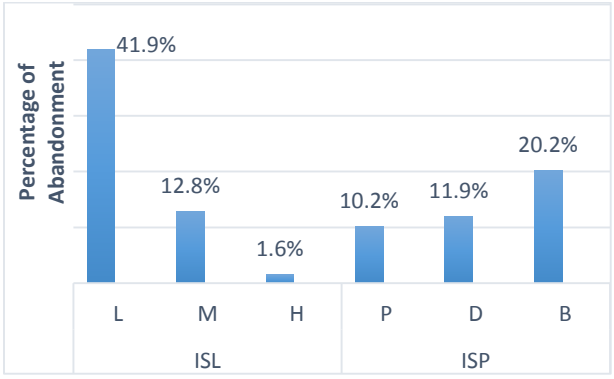


Figure 6. Abandonment by experimental condition (L=low; M=medium; H=high; P=persistent; D=disrupted; B=bursting).

Performance on each continuous search behavior measure given each ISL and ISP condition is reported in Tables 3 and 4. From Table 3 one can see as ISL increased from low to high, participants consistently spent more time searching in the search result set, examined more results, went to greater depths both in terms of clicks and mouse hovers, saved more documents, clicked but decided not to save more documents, and scanned through more non-relevant snippets. However, there are no obvious consistent differences among ISP condition (Table 4).

To examine whether the results observed from the descriptive statistics were significant**,** Generalized Estimating Equations (GEE) [41] was applied to model the effect of ISL, ISP and NFC on search behaviors. GEE allowed us to run linear regression analysis (for continuous measures) and logistic regression analysis (for categorical measures) for repeated measurements. For each search behavior measure, we entered ISL, NFC and their interaction term in one model, and ISP, NFC and their interaction term in another model. Results of GEE are reported in Tables 5 and 6 and are discussed in detail in the subsequent sections according to research question.

Table 3. Search behavior measures (*M, SD*) by ISL (The highest mean for each measure is bolded to facilitate comparisons)

Measures	Low	Medium	High
Time	61.85 (63.12)	101.86 (75.35)	**128.05 (79.28)**
NumPagination	.48 (1.27)	.47 (1.38)	**.61 (1.58)**
NumExamined	.78 (0.81)	2.02 (1.33)	**2.82 (1.37)**
DeepestRankClick	2.71 (5.45)	4.10 (4.70)	**6.49 (10.03)**
DeepestRankHover	9.23 (14.57)	9.85 (16.24)	**12.02 (18.51)**
NumPred	.21 (.47)	.48 (.70)	**.65 (.78)**
NumRele	.57 (.55)	1.47 (1.14)	**2.13 (1.28)**
NumNonRele	1.94 (4.87)	2.11 (4.09)	**3.67 (9.22)**

Table 4. Search behavior measures (*M, SD*) by ISP (The highest mean for each measure is bolded to facilitate comparisons)

Measure	Persistent	Disrupted	Bursting
Time	**121.17 (102.11)**	112.54 (76.39)	97.48 (77.16)
NumPagination	.47 (1.21)	.48 (1.45)	**.53 (1.45)**
NumExamined	**2.35 (1.69)**	2.31 (1.62)	1.95 (1.59)
DeepestRankClick	**6.33 (5.71)**	4.82 (5.21)	5.72 (4.34)
DeepestRankHover	10.74 (13.21)	10.05 (15.86)	**11.56 (16.79)**
NumPred	**.40 (.69)**	.37 (.58)	**.40 (.62)**
NumRele	**1.94 (1.61)**	1.91 (1.58)	1.49 (1.39)
NumNonRele	**3.98 (4.37)**	2.51 (4.33)	3.77 (3.26)

5.2 RQ1: ISL and Search Behaviors

The first research question of this study is: what is the relationship between ISL and search behavior? The results from Table 5 indicate that ISL significantly influenced Time, QueryAction, Abandonment, NumExamined, NumPred, NumRele, NumNonRele and DeepestRankClick, Follow-up contrasts were conducted to compare whether the differences between any two conditions were significant. The results show that high ISL led to the greatest NumExamined, DeepestRankClick, NumPred, and NumRele, followed by medium, and low (all contrasts: p <.05); high ISL led to more NumNonRele than low ISL (p <.05), but no significant difference was found between high and medium for NumNonRele. Participants abandoned their queries more often when the ISL was low, followed by medium, and high (all contrasts: p<.01). There were interaction effects between ISL and NFC on Time and QueryAction; all interaction effects are discussed in Section 5.5.

In most tasks, participants did not paginate to the second page, so we examined how ISL influenced interactions on the first SERPs

by only including cases where participants did not paginate. We found one additional significant effect of ISL: DeepestRankHover (X^2=20.08, p<.0001) All other measures remained statistically significant. When the first SERP had high ISL (M=4.68, SD=.20) participants hovered to lower ranks than medium (M=4.30, SD=.20) and low (M=3.39, SD=.30) (L<M, L<H; p<.01).

Table 5. Results for ISL conditions (Wald X^2, significance)

Measures	ISL	NFC	Interaction
Time	62.94****	9.15**	6.01*
NumPagination	0.74	3.99*	2.33
QueryAction	17.69**	0.20	13.17*
Abandonment	46.42****	43.56	0.44
NumExamined	219.26****	0.45	0.12
NumPred	33.90****	0.00	0.61
NumRele	110.85****	0.31	0.24
NumNonRele	6.47*	1.41	0.96
DeepestRankClick	30.33****	1.28	0.89
DeepestRankHover	2.02	4.78*	2.77

Note. *p<.05, **p<.01, ***p<.001, ****p<.0001

5.3 RQ2: ISP and Search Behaviors

The second research question addresses the relationship between ISP and search behaviors. Results in Table 6 show ISP had a significant effect on NumRele and NumNonRele. Follow-up contrasts indicate NumRele was significantly greater in the persistent and disrupted conditions than in bursting (p<.05), but no difference was found between persistent and disrupted. NumNonRele was significantly higher in persistent and bursting than in disrupted (p<.05), but no significant difference was found between persistent and bursting. Results also revealed that abandonment rates for persistent and disrupted were significantly lower than bursting (p<.05).

Table 6. Results for ISP conditions (Wald X^2, significance)

Measure	ISP	NFC	Interaction
Time	3.80	0.50	0.59
NumPagination	0.16	2.24	1.51
QueryAction	2.74	1.81	3.16
Abandonment	6.33*	0.58	3.34
NumExamined	5.52	0.42	1.50
NumPred	0.33	0.50	2.43
NumRele	8.57*	0.83	2.57
NumNonRele	9.98**	0.85	2.18
DeepestRankClick	5.07	0.80	2.54
DeepestRankHover	1.60	2.54	1.78

Note. *p<.05, **p<.01

Next, we excluded the queries that led to paginations and examined the search behavior measures within the range of the first page. It was found that participants stopped clicking at higher ranks in the disrupted condition (X^2=79.30, p<.01). While in both persistent and bursting conditions DeepestRankClick were similar (4.51 and 4.61, respectively), when the disrupted condition was encountered, participants did not click beyond rank 2.76. There were also significant effects of ISP on DeepestRankHover (P, B>D, p<.01), and NumNonRele (P, B>D, p<.05).

5.4 RQ3: NFC and Search Behaviors

The last research question examines the relationship between NFC and search behaviors. The effects of NFC according to ISL and ISP are reported in Tables 5 and 6. There was a main effect of NFC on Time, NumPagination, and DeepestRankHover in ISL conditions but not in ISP conditions. People with higher NFC scores paginated less and stopped hovering at higher ranks. The effect of NFC on time was conditioned on ISL and is discussed in the next section.

5.5 Interaction Effects

An interaction effect between ISL and NFC was found for time (X^2=6.01, p=.05). The relationships are plotted in Figure 7 at NFC=10th, 50th and 90th percentile for better understanding. While overall ISL was related positively to time and NFC was related negatively to time, the effect of ISL on time was moderated by NFC to a greater extent when ISL was high. As NFC increased from the 10th percentile to the 90th, the difference between ISL became less obvious. In other words, while all participants searched longer when there were more relevant documents, low NFC participants increased their search time more across ISL.

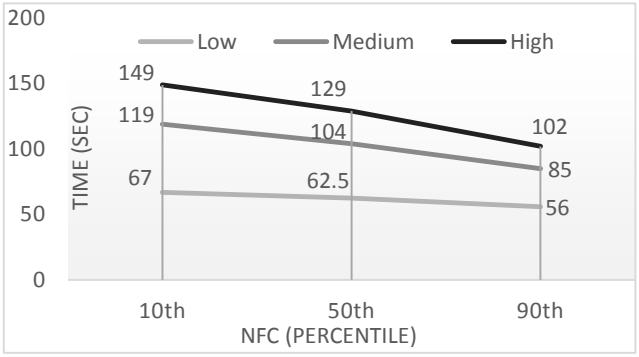


Figure 7. Interaction effect between ISL and NFC on Time

Another interaction effect between ISL and NFC was found in QueryAction (X^2=17.69, p<.01), which means the relationship between NFC and the predicted probability for each QueryAction outcome depended on ISL, and the nature of the effect of ISL varied with the outcomes. The predicted probabilities for reformulation, pagination and stopping are plotted in Figure 8. The relationship

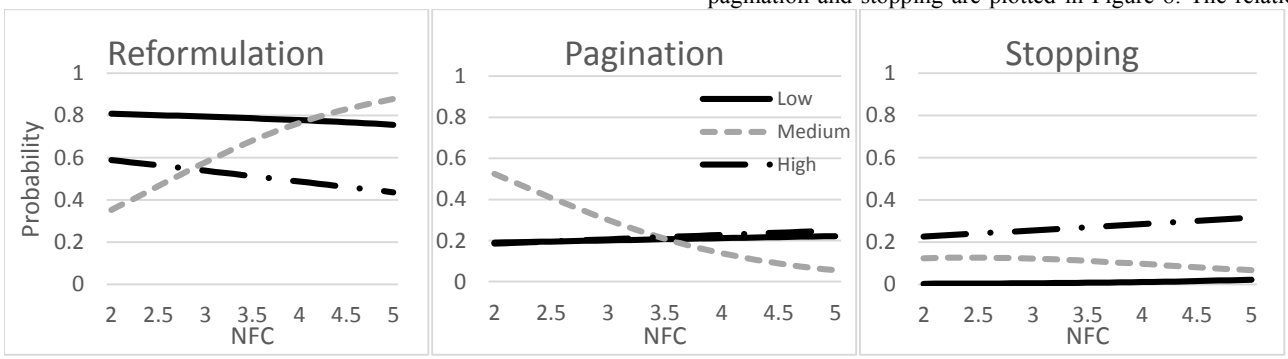


Figure 8. (Left to right) Predicted probability of reformulation, pagination, and stopping by NFC and ISL

between NFC and the probability of reformulation was positive for medium ISL, but negative for low and high ISL, indicating that higher NFC was related to a higher probability of reformulation for medium ISL but lower probabilities for low and high ISL. The slope is also the steepest for medium, which means the effect of NFC was the strongest for medium ISL. The relationship between NFC and pagination was negative and strong for medium but positive and weak for low and high ISL. Lastly, the relationship between NFC and stopping was weak and negative for medium, but weak and positive for low and high.

5.6 Participant Explanations of Behaviors

After participants finished their searches, they were shown video-recordings of three of their searches and asked to provide explanations whenever they reformulated, paginated or stopped.

5.6.1 Reformulation

Participants who reformulated on the first SERP often did so because the results suggested to them they needed to be more specific or they were not on the right track. For example, one participant pointed to the screen and commented "From here down, there is nothing to do with what I am looking for, something is telling me I need to re-strategize." This suggests that reformulations could result from lack of information scent. When prompted why they did not paginate, some people said they were so used to searching with Google that they believed relevant results should be on the first page, which probably explains why only 21% queries in our study led to pagination; even when participants knew relevant pages could occur after the first SERP, they found "it is easier to do a new search." Others said they just "did not bother to go through more results;" instead they wanted to find the right search word to bring good results on top. One participant acknowledged that it was the knowledge that "this isn't your last chance, right? I can always come back to this search configuration if I want to….it is a desire, a sense that a new term might get me better results" encouraged him to re-query instead of going to the second page. Others issued another query because they had acquired enough information about one aspect of the topic, and decided to shift focus.

5.6.2 Paginations

Participants often attributed paginations to the observation that the first page looked good but they still did not have enough information. Others said they continued because they were curious about what was on the second page, they did not know what other terms to use, or they felt confident about their search terms. For participants who mentioned the first page looked good, they often based quality judgments on "the number of relevant results" or "the aggregate quality" observed on the first page; hardly did they discuss the effect of ranking on pagination. When participants paginated, they were asked why they did not choose to submit a new search instead. Some participants believed going to the second page was "just a click, why not?" Compared with participants who preferred to invest cognitive effort in query formulations, this behavior appeared in a sharp contrast about how participants conceptualize search effort. Others expressed that learning from experience they knew sometimes good results were not on the first page because "not many people searched for them." Still others said they almost always gave the second and even the third page a chance out of habit, exemplifying a case where personality may come into play.

At the end of the interview, participants were encouraged to recall in their real life when they actually paginated to gather information. The most commonly described searches where participants paginated were when they conducted people search, product search, image search and literature search. In people search, participants commented they knew little about a person and many people could share the same names, so they did not mind filtering through search results. In product search, because they did not want to miss great deals, participants were willing to go deeper just in case. Participants said they paginated in image search because processing images was perceived less effortful than text. And in literature search some participants commented they often used Google Scholar or library databases, which they believed were more trustworthy, and therefore they were comfortable going through multiple pages, assuming results deeper were also credible.

5.6.3 Search Stopping

The strategies participants used to decide when to stop included fulfilling the task requirements, receiving both positive and negative perspectives, finding information from several credible sources, feeling they knew enough to have a discussion with the hypothetical family member in the task description, feeling they had spent enough time searching, and not feeling interested in the topic. A couple of participants used a pre-determined number of relevant documents as a heuristic for when to stop searching; these pre-determined values were 3 or 4 for some, and 5 or 6 for others. These comments corresponded to [31] where the researchers proposed searchers use a pre-determined threshold to decide when to stop. Few participants said they based their stopping decisions on "intuitions" or "feelings," as many previous studies have reported [15][34][44].

6. DISCUSSION

The current study examined the extent to which the theories of Information Scent and Need for Cognition could be used to explain three search behaviors: reformulation, search depth and stopping. The findings suggest that ISL was a stronger predictor than ISP. Not only did ISL predict search depth on the first SERP, together with NFC, they were able to explain the patterns of variations in the predicted probabilities of reformulation, pagination and stopping.

From Figure 8 we can see the slopes of medium ISL were steeper than the slopes of low and high when predicting the probabilities of reformulation and pagination, which means that NFC had a more profound effect when the SERP contained a mid-level amount of relevant result pages. The consistently higher reformulation probabilities in low ISL than high ISL regardless of NFC probably suggests that five relevant documents were adequate while one relevant document was not enough. Yet when there were three relevant documents, participants with higher and lower NFC possibly had different interpretations of whether this was enough. For participants with high NFC, evaluating three relevant results probably reached their higher bound of "the feeling of enough" for a given query submission, therefore the probability of reformulation increased and the probability of pagination decreased. On the other hand, for participants with low NFC, examining three relevant documents probably only reached their lower bound of "the feeling of enough", which explained why there was a lower probability of reformulation and a higher probability of pagination.

Even though ISP effects were not obvious in this study, at least the alternative rankings manipulated in this study, persistent and bursting ISP, did not perform worse than disrupted ISP, the ranking which was modeled after today's search engine result ranking. Greater DeepestRankClick and DeepestRankHover on the first SERP in persistent and bursting suggested that the similar performances were due to extra search effort on the first SERP. This finding repeated previous findings that searchers work harder to compensate for less than optimal result rankings [38]. This

finding implies that during the testing of different search rankings that people are tolerant of some non-relevant results so long as they are positioned at specific ranks. A possible implication in the context of aggregated search is that when search engines are not confident about which vertical results (e.g., image or news) to present, interleaving vertical results of questionable relevance with relevant text results may cause users to stay on the SERP. Given an ambiguous query (e.g., mac), interleaving Web results and vertical results of different senses (e.g., Mac the computer and MAC the makeup) might engage users and allow for search engines to gather more user interaction information.

ISL and ISP impact on query abandonment demonstrates the usefulness of information scent on information source selection. When the first SERP showed only one relevant result, participants probably assumed the results were not promising just by the snippets and abandoned the results set; our participants' comments that a low number of relevant pages indicated they were "not on the right track" supported this explanation. When participants were presented with the bursting ISP, it was possible participants did not realize there were four relevant documents since they did not scan deep enough. Persistent ISP could have maintained participants' attention because the interleaving of relevant and non-relevant results prompted them to reflect on their queries. It might also be the case that as long as the first two results were relevant, the information scent provided by the SERP was enough to engage participants.

The high reformulation rate observed in the current study suggests a reconsideration of the use of search effectiveness measures. While researchers often decide an arbitrary rank at which users may reach based on hypothetical user models and task models, setting a rank beyond 10 seems unrealistic in most cases, even for information-gathering tasks. Our results show that even in open-ended tasks where participants needed more than one result (or snippet) to solve an information problem, participants tended to reformulate to gather more information than paginate. Even in high ISL where there were five relevant results at the optimal ranking, the reformulation rate was still as high as 50%. If most results beyond the 10th rank are never examined by people when using search engines, evaluating algorithms based on an inclusion of results beyond the 10th rank probably does not reflect users' perception of system performance at least for the types of tasks typically evaluated in experimental IR. Note, however, that our participants did identify several scenarios (e.g., database search) and tasks (e.g., people search) where they typically paginated.

Differences in NFC provided a strong explanation for search depth. Results showed that participants with higher NFC paginated less and had more shallow hovers. Our results also showed that lower NFC participants went deeper in the search result. This implies that search results that occupy the highest ranks play a more important role in search result evaluation for high NFC searchers than for low NFC searchers (results ranked lower were less likely to be viewed by high NFC searchers). Thus, a different weighting scheme or discounting factor in measures such as nDCG might be used based on the searcher's NFC. While higher NFC participants remained relatively stable in terms of the time they spent on search result sets regardless of ISL, participants with lower NFC showed greater adaptability to the SERP environment and were more willing to go deeper in the SERP. With respect to interface design, these results suggest that instead of displaying ten search results per page, displaying more results per page or automatically loading subsequent pages as participants scroll, might encourage low NFC searchers to consider more information before making decisions. In situations where searchers have higher motivation to paginate than to reformulate, slightly different interventions can be applied.

7. CONCLUSIONS

In this work we investigated the extent to which two theories, Information Scent and Need for Cognition, explained search behaviors. The findings showed that the *level* of information scent on the first SERP and a searcher's need for cognition can be used to explain search behaviors such as reformulations, search depth and stopping. Although this study sought to evaluate the usefulness of two theories for explaining search behavior, clearly there is more work to be done if we are to establish more explanatory models of search behavior to complement the numerous descriptive and predictive models that currently exist.

8. ACKNOWLEDGEMENT

We thank Drs. Jaime Arguello, Nicholas Belkin, Robert Capra and Barbara Wildemuth for their input about research design, and Dr. Chris Wiesen for his assistance with statistical analysis.

9. REFERENCES

[1] Anick, P. (2003). Using terminological feedback for Web search refinement - A log-based Study. In *Proceedings of SIGIR'03,* 88–95.

[2] Azzopardi, L., Kelly, D., & Brennan, K. (2013). How query cost affects search behavior. In *Proceedings of SIGIR'13, 23–32.*

[3] Bailey, J. (1997). Need for cognition and response mode in the active construction of an information domain. *Journal of Economic Psychology*, *18*(1), 69–85.

[4] Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3), 152.

[5] Browne, G., & Pitts, M. (2004). Stopping rule use during information search in design problems. *Organizational Behavior and Human Decision Processes, 95*(2), 208–224.

[6] Cacioppo, J., Petty, R., Feinstein, J., & Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Review*, *119*(2), 197–253.

[7] Cacioppo, J., Petty, R., & Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.

[8] Cacioppo, J., Petty, R., & Morris, K. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology, 45*, 805–818.

[9] Card, S., Pirolli, P., Van Der Wege, M. M., Morrison, J. B., Reeder, R. W., Shraedley, P. K., & Boshart, J. (2001). Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability. In *Proceedings of CHI,* 498-505.

[10] Chi, E. H, Pirolli, P., Chen, K., & Pitkow, J. E. (2001). Using information scent to model searcher needs and actions on the web. In *Proceeding of CHI, (CHI Letters)*, *3*(1), 490–497.

[11] Chi, E. H, Pirolli, P., & Pitkow, J. (2000). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a Web site. In *Proceeding of CHI,* 161–168.

[12] Cooper, W. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak

ordering action of retrieval systems. *American Documentation, 19*(1), 30–41.

[13] Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness part II. *JASIS, 24*(6), 413–424.

[14] Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in web search. In *Proceeding of CHI,* 407–416.

[15] Dostert, M., & Kelly, D. (2009). Searchers' stopping behaviors and estimates of recall. In *Proceedings of SIGIR*, 820–821.

[16] Diriye, A., White, R. W., Buscher, G., & Dumais, S. T. (2012). Leaving so soon? Understanding and predicting Web search abandonment rationales. In *Proceedings of CIKM,* 1025-1034.

[17] Ford, N., Miller, D., & Moss, N. (2001). The role of individual differences in Internet searching: An empirical study. *JASIS&T*, *52*(12), 1049–1066.

[18] Hassan, A., Shi, X., Craswell, N., & Ramsey, B. (2013). Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proceedings of CIKM,* 2019-2018.

[19] Huang, J., & Efthimiadis, E. N. (2009). Analyzing and evaluating query reformulation strategies in web search logs. In *Proceeding of CIKM,* 77-86.

[20] Huang, J., White, R. W., & Dumais, S. (2011). No clicks, no problem. In *Proceeding of CHI*, 1225–1234.

[21] Jansen, B. J., Booth, D. L., & Spink, A. (2009). Patterns of query reformulation during Web searching, *JASIS&T*, *60*(7), 1358–1371.

[22] Jansen, B.J., & Spink, A. (2005). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *IP&M, 42*(1), 248–263.

[23] Järvelin. K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM TOIS*, 20(4), 422–446.

[24] Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*, 154–161.

[25] Kammerer, Y., Wollny, E., Gerjets, P., & Scheiter, K. (2009). How authority-related epistemological beliefs and salience of source information influence the evaluation of Web search results–An eye tracking study. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2158-2163.

[26] Kantor, P. B. (1987). A model for the stopping behavior of searchers of online systems. *JASIS, 38*(1), 211–214.

[27] Kraft, D., & Lee, T. (1979). Stopping rules and their effect on expected search length. *Information Processing and Management*, *15*(1), 47–58.

[28] Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., & Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *JASIS&T*, *59*(7), 1041–1052.

[29] Loumakis, F., Stumpf, S., & Grayson, D. (2011). This image smells good: Effects of image information scent in search engine results pages. In *Proceedings of CIKM, 11,* 475–484.

[30] Mansourian, Y., & Ford, N. (2007). Search persistence and failure on the web: a "bounded rationality" and "satisficing" analysis. *Journal of Documentation*, *63*(5), 680–701.

[31] Nickles, K. R., Curley, S. P., & Benson, P. G. (1995). Judgment-based and reasoning-based stopping rules in decision making under uncertainty. *Working Paper, Wake Forest University.*

[32] Nov, O., Arazy, O., López, C., & Brusilovsky, P. (2013). Exploring personality-targeted UI design in online social participation systems. In *Proceedings of CHI*, 361-370

[33] Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review,* 106(4), 643–675.

[34] Prabha, C., Connaway, L. S., Olszewski, L., & Jenkins, L. R. (2007). What is enough? Satisficing information needs. *Journal of Documentation*, *63*(1), 74–89.

[35] Rieh, S. (2002). Judgment of information quality and cognitive authority in the Web. *JASIS&T, 53*(2), 145–161.

[36] Rieh, S.Y. and Xie, H. (2006). Analysis of multiple query reformulations on the web: the interactive information retrieval context. *IP&M, 42*(3), 751-768.

[37] Scholer, F., Kelly, D., Wu, W. C., Lee, H. S., & Webber, W. (2013). The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of SIGIR,* 623-632.

[38] Smith, C. L., Kantor, P. B., & Street, H. (2008). User adaptation : Good results from poor systems. In *Proceedings SIGIR,* 144–154.

[39] Spink, A., Jansen, B. J., & Ozmultu, H. C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research*, *10*(4), 317–328.

[40] Toms, E., & Freund, L. (2009). Predicting stopping behaviour: A preliminary analysis. In *Proceedings of SIGIR,* 750–751.

[41] Vittinghoff, E. (2012). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. Boston: Springer.

[42] White, R. W., & Dumais, S. T. (2009). Characterizing and predicting search engine switching behavior. In *Proceedings of CIKM,* 87-96.

[43] Wu, W.C., Kelly, D., Edwards, A. & Arguello, J. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. *Poster presented in the 4th Information Interaction in Context Symposium (IIiX'2012)*.

[44] Zach, L. (2005). When is "enough" enough? Modeling the information-seeking and stopping behavior of senior arts administrators. *JASIS&T, 56*(1), 23–35.