Task Complexity, Vertical Display and User Interaction in Aggregated Search

Jaime Arguello, Wan-Ching Wu, Diane Kelly, Ashlee Edwards
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599-3360 USA
{jarguello, wanchinw, diane.kelly, aedwards}@unc.edu

ABSTRACT

Aggregated search is the task of blending results from specialized search services or *verticals* into the Web search results. While many studies have focused on aggregated search techniques, few studies have tried to better understand how users interact with aggregated search results. This study investigates how task complexity and vertical display (the blending of vertical results into the web results) affect the use of vertical content. Twenty-nine subjects completed six search tasks of varying levels of task complexity using two aggregated search interfaces: one that blended vertical results into the web results and one that only provided indirect vertical access. Our results show that more complex tasks required significantly more interaction and that subjects completing these tasks examined more vertical results. While the amount of interaction was the same between interfaces, subjects clicked on more vertical results when these were blended into the web results. Our results also show an interaction between task complexity and vertical display; subjects clicked on more verticals when completing the more complex tasks with the interface that blended vertical results. Subjects' evaluations of the two interfaces were nearly identical, but when analyzed with respect to their interface preferences, we found a positive relationship between system evaluations and individual preferences. Subjects justified their preference using similar rationales and their comments illustrate how the display itself can influence judgments of information quality, especially in cases when the vertical results might not be relevant to the search task.

Categories and Subject Descriptors

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

General Terms

Performance, Experimentation, Human Factors.

Keywords

Aggregated search interfaces, search behaviors, evaluation, user study, interaction, task complexity

1. INTRODUCTION

In addition to Web search, commercial search companies (e.g., Google, Bing, Yahoo!) provide access to a wide range of specialized services known as *verticals* (e.g., images, video, news). There are two ways that users can access vertical results. If a user wants results from a particular vertical, the query can be issued directly to the vertical-specific search engine. In other cases, however, a user may not know that a vertical is relevant or may want results from multiple verticals at once. For these reasons, commercial systems often showcase vertical results alongside the Web results. Currently, this is done by blending a few of the vertical's results somewhere above, within, or below the Web results. The goal is to either satisfy the user with the blended results or to convey how the information need might be better satisfied by directly searching the vertical. The task of surfacing vertical results in response to a Web search query is known as *aggregated search*.

Most published research in aggregated search has focused on automatic methods for predicting which verticals to present (*vertical selection*) [4, 5, 11, 19] and where in the Web results to present them (*vertical presentation*) [2, 3, 23]. Evaluation of these systems has typically been conducted by using editorial vertical relevance judgements as the gold standard [2, 3, 4, 5, 19], or by using user-generated clicks on vertical results as a proxy for relevance [11, 23]. While these studies have greatly advanced the state of the art in aggregated search techniques, because users are far removed from the evaluation, they have contributed little insight about how users' higher-level objectives influence their engagement with vertical search results.

A few published studies have investigated user behavior with aggregated search interfaces [24, 25, 28]. Thus far, these studies show two major trends. First, when a vertical is relevant, users prefer to see its results towards to the top of the blended results [25, 28]. Second, there seems to be a click-bias in favor of visually appealing verticals such as *video* [25]. While these studies reveal important trends, several questions remain. Prior work shows that when completing more complex tasks users take longer, enter more queries, view more results, and use more sources than when completing less complex tasks [20]. Are these same effects *also* observed with aggregated search interfaces? What is the effect of task complexity on a user's demand for vertical results? Are more vertical results examined when the task is more complex? And, does it depend on whether the vertical results are blended into the web results? This study investigates these questions.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA. Copyright 2012 ACM 978-1-4503-1472-5/12/08...$15.00.

1.1. Research Questions

This study investigates how users interact with interfaces that vary the way that vertical results are displayed in response to a query. We investigated two different interfaces, referred to as the *blended interface* and the *non-blended interface.* Given the same query, both interfaces provided *indirect* access to the same set of Web and vertical results. Navigational tabs across the top and left side of both interfaces provided links to issue the query (or a new query) to only the Web search engine or only a particular vertical search engine. The only difference was that the blended interface showcased a few results from every vertical in the main search results page, while the non-blended interface did not.

In this work, we also focus on tasks of varying level of cognitive complexity and the effect these tasks have on users' interactions with vertical results *and* whether the amount of interaction depends on the blending of vertical results into the Web results. Prior work shows that when completing more complex tasks users take longer, enter more queries, view more search results, and use more sources than when completing less complex tasks [20]. However, no work has examined task complexity in the context of aggregated search. Specifically, we address the research questions in Table 1.

Table 1. Research Questions and Hypotheses

Research Question	Hypothesis
RQ1: How does task complexity affect search interaction in aggregated search?	**H1**: People will interact more when conducting more complex tasks.
RQ2: How does task complexity affect use of vertical search results?	**H2**: People will use more vertical results for more complex tasks.
RQ3: How does vertical display affect use of vertical results?	**H3**: People will use more vertical results when these are blended into the Web results.
RQ4: What are people's evaluations of the interfaces, perceptions of vertical results, and display preferences?	This question is primarily explored with no hypothesis.

2. RELATED WORK

2.1. Aggregated Search

The goal of aggregated search is to blend results from zero or more verticals into the Web search results. The task is typically decomposed into two subtasks: predicting which verticals to present (*vertical selection*) and predicting where in the Web results to present them (*vertical presentation*). Existing methods for vertical selection and presentation use machine learning to combine different types of predictive evidence: query-string features [2, 4, 5, 19, 23], vertical query-log features [2, 4, 5, 11, 23], vertical content features [2, 4, 5, 11], and implicit feedback features from previous presentations of the vertical [11, 23]. Model tuning and evaluation is typically done with respect to editorial relevance judgements [2, 3, 4, 5, 19] or, in a production environment, with respect to user-generated clicks and skips [11, 23]. In the first case, users do not actively participate in the evaluation. In the second case, their feedback is not *explicit*. Because users are far removed from the evaluation, this research has contributed little insight about how properties of the higher-level task (e.g., its complexity) influence users' interaction with and evaluation of the system.

A few user studies have investigated search and preference behavior with aggregated search interfaces [24, 25, 28]. Sushmita *et al.* [24] experimented with two aggregated search interfaces: one similar to our non-blended interface, which did not combine content from different sources in the main results, and a second interface that combined content from different sources in a two-dimensional blocked fashion. Subjects were asked to complete search tasks using both systems. Results showed significantly more clicks on vertical results using the aggregated interface. However, differences in the *inclusion* of cross-vertical content in subjects' response sheets and differences in subjects' system preferences were not significant.

In later work, Sushmita *et al.* [25] conducted a similar study with an aggregated interface similar to our blended interface and a blocked interface similar to the one used in Sushmita *et al.* [24]. Results showed that users clicked more on verticals that were presented higher in the blended results and on verticals that were more relevant to the task. Furthermore, a click bias was observed in favor of *video* results, which were more visually salient. That is, compared to the other verticals, users clicked more on video results irrespective of rank and relevance.

While the work above focused on different aggregated search interfaces, Zhu and Carterette [28] focused on where to blend results from a single vertical: *images*. They found a preference in favor of images ranked high in the blended results for queries with image intent.

To summarize, the work above reveals a few important trends. Users click more on verticals that are relevant and prefer layouts where the relevant verticals are ranked higher in the blended results. Additionally, aggregated search is not immune to click bias. Irrespective of relevance, users click more on verticals that are ranked high and are visually salient. Our work extends this previous work by focusing on the effect of task complexity and vertical display (the blending of vertical results into the main search results) on the use of vertical results and on users' evaluations of the system. Thus, one important aspect of our work is the manipulation of task complexity as an independent variable.

2.2 Tasks and Task Complexity

Toms [26] (quoting Hackman [12]) defines a task as a "set of assigned (a) goals to be achieved, (b) instructions to be performed, or (c) a mix of the two" ([26], pg. 45). Li and Belkin [17] define an information search task as "a task that users need to accomplish through effective interaction with information systems" ([17], pg. 1823). Information *search* tasks are usually distinguished from *work* tasks [7]. In this study we focus on *search* tasks. While task has always been an essential part of interactive search studies, it has been increasingly used as an independent variable and many studies have demonstrated that search behavior varies according to task and task characteristics [8, 14, 15, 20].

A large body of research has attempted to conceptualize and define tasks and task characteristics (e.g., [7, 9, 17, 26, 27]). Researchers have classified tasks according to type (e.g., open, factual, navigational, decision-making) and according to task properties (e.g., difficulty, urgency, structure, stage). Li and Belkin [17] present a faceted classification of tasks. This classification includes *generic* facets of tasks (e.g., source of task, time, product, process and goal) and *common* facets of tasks including characteristics (e.g., objective task complexity and interdependence) as well as users' perception of task (e.g., salience, urgency, difficulty, subjective task complexity and knowledge of task topic). Li [16] later found that the facets of *product* and *objective task complexity* had the most significant impact on search behavior. In a follow-up study, Li and Belkin [18] found that objective task complexity affected users' search interactions such as number of queries issued, mean query length, number of pages viewed, and number of sources consulted. Liu *et al.* [20] also found that objective task complexity (as measured by activities and information required) impacted search behavior. When completing more complex tasks, their subjects took longer, entered more queries, viewed more pages, and used more sources.

In this paper, we use cognitive complexity as a way to model differences between search tasks. We used three conceptualizations of task complexity to guide the creation of our search tasks (presented in Section 3.3). The first was described by Campbell [9] and later used by Li and Belkin [17]. In this model, four attributes determine complexity: (1) the number of potential paths to the desired outcome; (2) the presence of multiple desired outcomes; (3) the presence of conflicting interdependencies between paths; and (4) uncertainty regarding paths.

The second conceptualization is from Byström and Järvelin [8], who define task complexity as the *a priori determinability* of tasks, which is the extent to which the searcher can deduce the required task inputs, processes, and outcomes based on the initial task statement. Although this conceptualization is based on subjective task complexity, Bell and Ruthven [6] used this model to create artificial search tasks with different levels of task complexity. The researchers reduced Byström and Järvelin's five-level categorization into a three-level categorization and found that objective task complexity was correlated with users' subjective assessments of complexity.

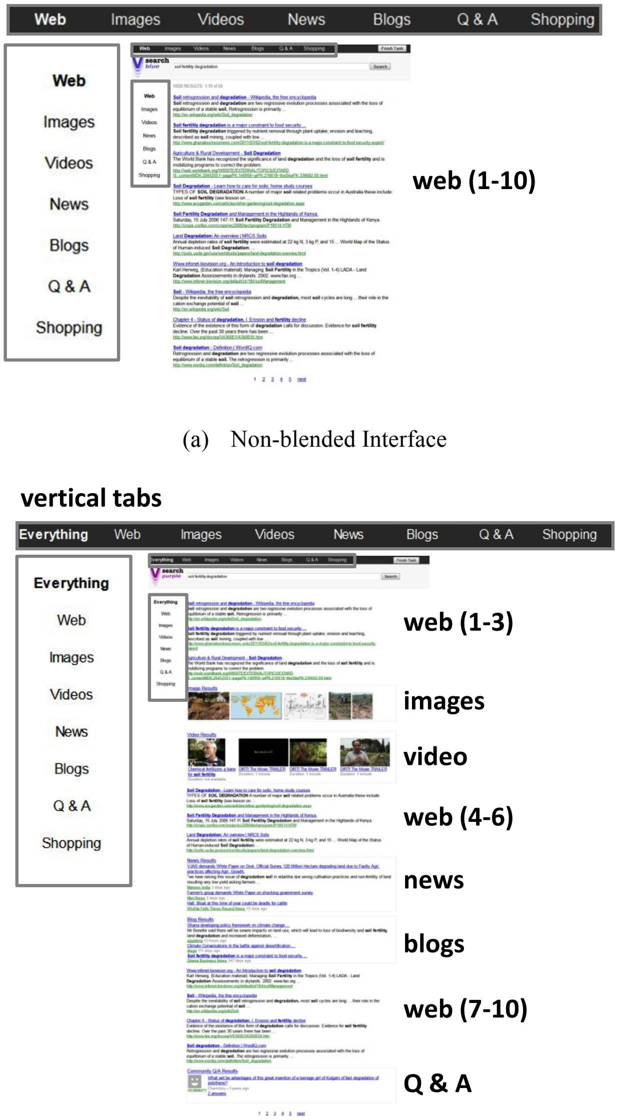
The third conceptualization of task complexity we use comes from Jansen *et al.* [13] who used Anderson and Krathwohl's taxonomy of educational objectives [1] to design search tasks. Anderson and Krathwohl's taxonomy of educational objectives has two dimensions: a cognitive process dimension and a knowledge dimension. Jansen *et al.* created tasks reflecting six types of cognitive processes: *remember*, *understand*, *apply*, *analyze*, *evaluate* and *create* (see Table 2 in Section 3.3 for definitions). Although Jansen *et al.* [13] did not situate this work in the context of task complexity, they observed a number of significant differences in the amount of interaction users exhibited when completing different task types, including session duration, number of queries, and number of pages viewed.

3. Method

The present study used a within subjects design. Subjects were randomly assigned to use the non-blended or blended interface first. Subjects completed three search tasks of three levels of task complexity on a single domain with each interface (described in more detail below). The order of task complexity levels was fixed for each subject across interfaces but was counterbalanced using a Latin Square design among subjects. Each subject conducted tasks from a single domain on each interface and the order of task domains was counterbalanced using a Latin square. Search sessions were logged using the Lemur Query-Log Toolbar.[1]

3.1. Interfaces

Two aggregated search interfaces were compared in this study: the non-blended interface and the blended interface (Figure 1). Both interfaces provided access to a Web search engine and six different vertical search engines: *images*, *videos*, *news*, *blogs*, *community Q&A*, and *shopping*. All search engines were constructed using freely available APIs from Bing (*web*, *images*, *video*, *news*), Google (*blogs*), Yahoo! (*community Q&A*), and eBay (*shopping*). Search results of each type were displayed similar to how they are displayed in commercial systems. News results were presented by displaying the article title, news source, and publication date. Shopping results were presented by displaying an image of the product, its name, and its price. Video results were presented by displaying an image of the video, its title, and duration.



(a) Non-blended Interface

(b) Blended Interface

Figure 1. Screenshots of the (a) non-blended and (b) blended interface with the top and left navigational bars enlarged. The shopping vertical is not shown in the blended interface because it did not retrieve results for this query.

[1] http://www.lemurproject.org/querylogtoolbar/

Both interfaces were identical with one exception. The non-blended interface did not showcase vertical results in the main search results page. To see results from a particular vertical in the non-blended interface, users had to click on the vertical tab located on the top or left navigation bar. This functionality was also available in the blended interface. However, the blended interface also showcased a few results from each vertical in the main search results page. Each set of vertical results included a link to the vertical search engine. In both interfaces, users could click on a vertical tab and issue subsequent queries directly to the vertical-specific search engine. In the blended interface, verticals were displayed in the same position relative to the Web results. While commercial systems vary the presentation of verticals depending on the query, we wanted to keep the user experience with the blended interface as consistent as possible.

The blended interface represents the baseline, as this is how vertical results are typically presented by commercial systems. However, we are not putting the non-blended interface forward as a novel way to present vertical results. Rather, it allows us to study the effects of vertical display on behaviors and preferences.

3.2. Subjects

In this study, we wanted to focus on populations that have not been studied with great frequency in laboratory studies of information search. To achieve this goal, we recruited participants from Pittsboro, North Carolina, a nearby rural township located approximately 20 miles from our university. We describe the community using data from the 2010 US Census to provide setting context, not because we claim that our sample is representative of this population. According to 2010 US Census data, Pittsboro, NC has a population of 3,743 and a population density of 1,100.9/sq mi. The median age of community members is 41.4. About 72% of community members identified themselves as white, 20% as black or African American, 1.8% as Asian and 3.7% as some other race. In addition, about 9% identified themselves as Hispanic or Latino. With respect to educational attainment, about 15% are high school drop-outs, 30% have a high school degree only, 22% have some college, but no degree, and 29% of the population has a bachelor's degree or higher. The median household income is $40,056 and about 18% of all people live below the poverty line. About 63% are in the workforce.

Twenty-two flyers were posted around the community at places such as laundromats, restaurants, grocery stores, convenience stores, and the public library. An email recruitment notice was sent to a community mailing list. The flyers and recruitment notice specified three inclusion criteria: at least 18 years of age, at least two years of online search experience, and proficiency in reading and writing in English.

We conducted our study in private rooms at the community public library and college. The community college and public library are on the same campus; the public library serves as the library for students at the community college. The first 8 respondents were used as pilot subjects; an additional 29 respondents were used in the actual study. Because of a logging failure, one of these subjects was excluded from analyses.

Our subjects consisted of 12 men and 16 women. The average age of our subjects was 42 years old, and the median age was 38. Our youngest subject was 20, and our oldest subject was 74 years old. Fifty percent of our subjects were professionals, and 17% worked in the service industry. Three of our subjects were students, 3 were unemployed, and 3 were retired. With respect to highest level of educational attainment: 18% of subjects earned a high school degree or GED, 11% an Associate degree, 53% a bachelor's degree, 14% a master's degree, and 4% a doctorate. All participants spoke English as a first language.

Eighty-six percent of our subjects said they had been using desktop or laptop computers for 10 or more years, 11% said 4-6 years, and 1 person said they had used computers for less than 1 year. The majority of subjects (96%) said they use computers daily and have regular access to computers. Search engine use among our subjects was varied. Ninety–six percent of subjects use Google, 64% use Yahoo Search, 46% use Bing, and 21% use AOL Search. The most common tasks our subjects reported performing online were searching information and accessing email, followed by browsing and surfing the Internet.

Subjects' search experience was measured with a modified version of the Search Self-Efficacy Scale [10], which contained 14 items describing different search activities. Subjects indicated their confidence in completing each activity using a 10-point scale, where 1=totally unconfident and 10=totally confident. Items were then averaged for each subject to arrive at a composite measure of Search Self-Efficacy. Subjects' average Search Self-Efficacy was 7.59 (SD=1.53). Because the scale was slightly modified from its original form, two internal consistency estimates of reliability were computed: Cronbach's alpha and Spearman-Brown split-half coefficient. Both coefficients were high (Cronbach's alpha=0.932 and Spearman-Brown=0.937), indicating strong internal consistency.

3.3. Tasks

In two recent studies of vertical search, Sushmita *et al.* [24, 25] investigated different types of tasks. In Sushmita *et al.* [24] users were asked to complete non-navigational tasks that required them to compile information from several sources. The tasks (according to the example provided in Sushmita *et al.* [24]) included an indicative request which suggested that pictures and news about recent events would be useful information to collect. Thus, these tasks implicated the use of verticals. In Sushmita *et al.* [25], the researchers created tasks that had different source orientations. For example, some tasks implicated videos, while others implicated images. In this study, we wanted to create tasks that might require verticals, but we did not want to tell subjects to look for these types of results. We decided instead to construct tasks that we believed would require different amounts and diversity of information to complete and different amounts of search effort. In doing so, we turned to Anderson and Krathwohl's taxonomy of educational objectives [1] and the work of Jansen *et al.* [13].

We focus on the *cognitive process* dimension of Anderson and Krathwohl's taxonomy, which is presented in Table 2. Six types of cognitive processes are identified: *remember*, *understand*, *apply*, *analyze*, *evaluate* and *create*, with each type requiring increasing amounts of cognition and effort. While this taxonomy is traditionally used to create educational materials such as exercises and exam questions, we used it to construct search tasks similar to Jansen *et al.* [13]. We propose that the different dimensions in Table 2 reflect increasing levels of cognitive complexity. Our definition of complexity is similar to that proposed by Li and Belkin [17] and Campbell [9] who related objective task complexity to the number of possible paths that can be taken to solve a task, the number of possible solutions and outcomes, and the amount of uncertainty. Since we are using learning tasks, we also propose that task complexity is related to the amount of cognition and effort required to complete the tasks.

Table 2. Anderson and Krathwohl's Taxonomy of Learning Objectives (Cognitive Process Dimension) [1]

Dimension	Definition
Remember	Retrieving, recognizing, and recalling relevant knowledge from long-term memory.
Understand	Constructing meaning from oral, written, and graphic messages through interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining.
Apply	Carrying out or using a procedure through executing, or implementing.
Analyze	Breaking material into constituent parts, determining how the parts relate to one another and to an overall structure or purpose through differentiating, organizing, and attributing.
Evaluate	Making judgments based on criteria and standards through checking and critiquing.
Create	Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing.

For our tasks, we used all the cognitive processes except *apply* because we were unable to create search tasks for this category that were distinct from the other categories. We selected five domains to use when creating the tasks: health, e-commerce, entertainment, travel and science. In total, we created 25 tasks (one for each cognitive process/domain combination). Pilot tests with 6 people allowed us to refine the tasks. Ultimately, in this study we only used tasks that required three types of cognition: *remember*, *understand* and *analyze*. This decision was made because our pilot subjects could not complete all the tasks during the 1.5 hour time allotment. Thus, in total we used 15 search tasks representing three levels of complexity and five domains. Example tasks from two domains are displayed in Table 3.

Table 3. Example search tasks from two domains

Type	Complexity	Health	Science
Remember	Low	How many people in the U.S. are currently living with HIV?	What is the name of the deepest point in the ocean?
Understand	Medium	What are some long-term health risks faced by professional football players?	What are some human activities that can degrade soil fertility?
Analyze	High	What are some of the different types of artificial tanning methods? What health risks are associated with them?	You recently heard that a lot of trash ends up in the ocean. In which oceans or areas does most trash end-up and why?

3.4. User Experience Questionnaire

The sub-scales from O'Brien's Engagement Scale [21, 22] formed the basis of the sub-scales we used in this study to elicit subjects' system evaluations. The Engagement Scale [22] consists of a 31-item scale with 6 sub-scales which measure the following aspects of engagement (number in parenthesis indicates how many items are on each sub-scale): Focused Attention (7), Perceived Usability (8), Endurability (5), Novelty (3), Aesthetics (5) and Felt Involvement (3). For all scales, subjects respond by indicating their level of agreement with the items (1=strongly disagree; 5=strongly agree). Because the Engagement Scale was developed in the context of e-commerce applications, we reviewed the items and made the following modifications so that the sub-scales were better-suited for general search system evaluation: (1) replaced words such as "shopping" and "website" with "searching" and "system," respectively; (2) eliminated the *aesthetics* sub-scale because the general properties of our interfaces (e.g., color, fonts) were identical; (3) added a 5-item *Search Effectiveness* sub-scale because none of the original items were designed to evaluate vertical search and the quality of the search results; and (4) deleted one item from the *Focused Attention* sub-scale after three pilot subjects commented on its awkwardness and one item from the *Endurability* sub-scale that did not make sense given the study situation. Ultimately, we used the following sub-scales: Focused Attention, Felt Involvement, Perceived Usability, Endurability, and Search Effectiveness. The Search Effectiveness sub-scale consisted of the items in Table 4. Readers are referred to O'Brien [21] for the content of the other sub-scales. Reliability coefficients for all sub-scales are displayed in Table 5 along with the number of items for each sub-scale.

Table 4. Search Effectiveness Sub-Scale

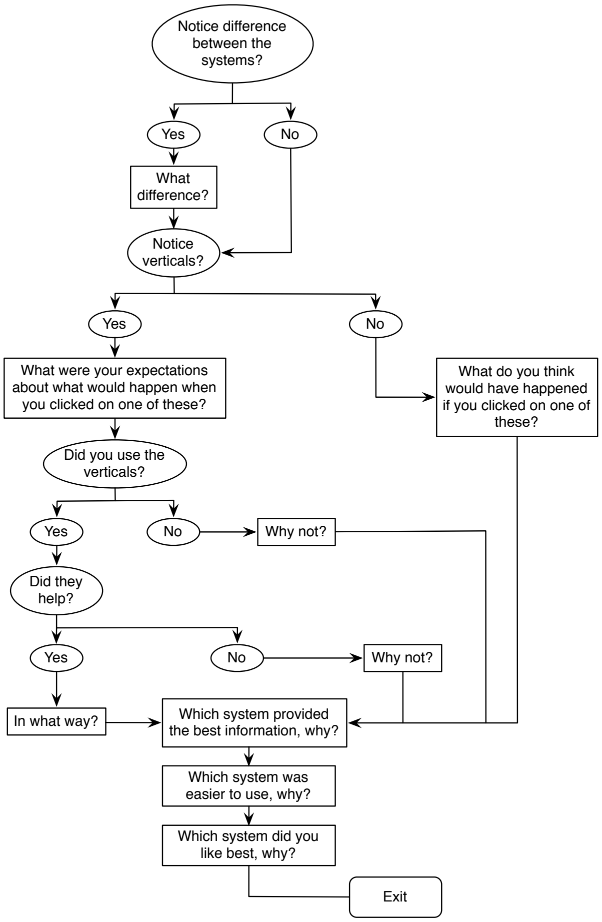
No.	Item
1	The system provided enough information to help me solve the search tasks.
2	The system provided me with many different kinds of information.
3	The presentation of search results helped me easily combine different types of information.
4	The presentation of search results allowed me to easily identify relevant information.
5	The presentation of search results helped me get an overview of the types of information available.

Table 5. Reliability Coefficients for Sub-scales

Sub-Scale	No. of Items	Cronbach's Alpha
Focused Attention	4	.789
Felt Involvement	3	.714
Perceived Usability	7	.940
Endurability	4	.802
Search Effectiveness	5	.864

3.5. Exit Questionnaire

The Exit Questionnaire (Figure 2) contained a number of open and closed questions and used display logic to customize the questions depending on how subjects responded. Subjects were first asked if they noticed any differences between the two systems. Those who responded *Yes* were asked to describe the differences they noticed and then went to the next question. Those who responded *No* skipped the differences question and went to the next question which displayed a screen shot of the basic interface (with no results displayed) and identified the various options along the top and side as "verticals." Subjects were first asked if they noticed these verticals. Both those who responded *Yes* and *No* were then asked similar questions about their expectations about clicking on the verticals. Those who responded *No* initially were then routed to the last three questions. Those who responded *Yes* were first asked if they used the verticals. Those who responded *No* were asked why not, while those who responded *Yes* were asked if the verticals helped them during their searches. Subjects responding *Yes* were asked to describe how they helped. Subjects were then routed to the last three questions. The last three preference questions asked subjects to indicate which system (1) provided the best information, (2) was easier to use, and (3) they liked best. Subjects were provided with a no difference option and asked to explain their choices.



Notice difference between the systems?
Yes
No
What difference?
Notice verticals?
Yes
No
What were your expectations about what would happen when you clicked on one of these?
What do you think would have happened if you clicked on one of these?
Did you use the verticals?
Yes
No
Why not?
Did they help?
Yes
No
Why not?
In what way?
Which system provided the best information, why?
Which system was easier to use, why?
Which system did you like best, why?
Exit

Figure 2. Exit Questionnaire Flow Diagram

3.6. Procedure

Once arriving to the study location, subjects were given an information sheet describing their rights as subjects and the basic procedure of the study. Subjects then completed an online entry questionnaire that collected information about demographics and computer and search experience. Next, subjects were given a set of written instructions. Following this, subjects completed a practice task using Google. The protocol required subjects to create responses to each search task in an auxiliary Word document. Subjects could type their response or use copy and paste. The purpose of the practice task was to acquaint subjects with this aspect of the study. Following the practice task, subjects were given their first search task and directed to use one of the interfaces. After completing three tasks, subjects completed the User Experience Questionnaire. Next, subjects repeated the entire sequence with another interface using different search tasks. Finally, subjects completed an Exit Questionnaire. The study took 1-1.5 hours to complete. Subjects were compensated $20 USD.

4. Results

We investigated four research questions: the effect of task complexity on search interaction (RQ1), the effect of task complexity on the use of vertical results (RQ2), the effect of vertical display on the use of vertical results (RQ3), and users' evaluations of each interface, their perceptions of verticals, and their interface preferences (RQ4).

4.1. Task Complexity & Interaction

With respect to RQ1, we hypothesized that more complex tasks would require a greater level of interaction (H1). While Anderson and Krathwohl's [1] taxonomy of learning tasks identifies six task-types, in this work, we focused on three (in ascending order of task complexity): *remember*, *understand*, and *analyze*. Search interaction was operationalized using five measures: time spent completing the task (time), number of queries issued (queries), number of search results clicked (clicks on SERP), and number of URLs visited (URLs). Subjects were asked to search naturally and, therefore, in several cases clicked on a search result and navigated far from the SERP. Thus, the number of URLs visited includes search results and URLs not directly linked to from the SERP. The number of SERP clicks included clicks on Web results as well as vertical results (in the blended interface).

Results showed a tendency for subjects to spend more time, issue more queries, click on more search results, and visit more URLs during tasks that were more complex (Table 6). A repeated measures ANOVA showed that all the differences were statistically significant.

Table 6. Interaction measures according to task complexity (Mean, St. Dev.)

	Time (sec)	Queries	Clicks on SERP	URLs
Remember	215.54 (185.26)	1.46 (0.83)	1.45 (1.09)	2.68 (3.61)
Understand	374.30 (221.53)	1.98 (1.51)	2.14 (1.73)	3.63 (3.15)
Analyze	491.93 (237.31)	2.88 (2.45)	2.79 (1.85)	4.75 (4.15)
F(2,110), *P*-value	39.85, *p*<.001	11.69, *p*<.001	14.57, *p*<.001	6.17, *p*=.003
Bonferroni post-hoc comparisons	A>U>R	A>U, R	A>U>R	A> R

Bonferroni post-hoc tests were used to pinpoint the differences and found that for time and clicks on SERP, the differences between all pairs of tasks were statistically significant in the manner hypothesized, with tasks of increasing complexity requiring increasing amounts of time to complete and SERP clicks. For Queries, subjects entered significantly more when completing the *analyze* tasks than *understand* and *remember* tasks. For URLs, the significant difference was between *analyze* and *remember* tasks. To summarize, this analysis supports our hypothesis (H1) that more complex tasks are associated with greater levels of search interaction.

4.2. Task Complexity, Vertical Display & Vertical Usage

With respect to RQ2 and RQ3, we hypothesized that subjects would click on more vertical results for more complex tasks (H2) and that subjects would click on more vertical results using the blended interface (H3). While we do not have a hypothesis for an interaction effect between task complexity and vertical display, we wanted to explore whether the effect of task complexity on vertical usage is contingent on vertical results being blended into the Web results. To examine these hypotheses, a 3 x 2 two-way ANOVA was conducted on vertical clicks. Results show a main effect for vertical display (*F*(1, 162)=10.25, *p*=.002), but no main effect for task complexity (*F*(2,162)=1.27, *p*=.282) and no significant interaction between vertical display and task complexity (*F*(2, 162)=2.61, *p*=.077).

As shown in Table 7, subjects clicked on significantly more vertical results when using the blended interface (M=.17, SD=.43) than when using the non-blended interface (M=.01, SD=.11). This, combined with the significant main effect described above for vertical display, confirms hypothesis (H3) that people use more vertical results when these are blended into the Web results.

It is worth noting that the number of results (Web and vertical results) presented in the blended interface's main results page was greater than the number of results (only Web results) presented in the non-blended interface's main results page. It could be that subjects clicked on more vertical results using the blended interface because they clicked on more results in general. However, based on a two-way ANOVA, the main effect of interface on all SERP clicks (Web and vertical) was not significant (*F*(1, 162)=.021, *p*=.885). Given a similar number of total clicks, subjects seem to prioritize clicking on vertical results in the blended interface.

Table 7. Vertical clicks and total SERP clicks across task complexity and interfaces. Mean (St. Dev.)

	Vertical clicks			SERP Clicks		
	Non-Blended (n=28)	Blended (n=28)	Total	Non-Blended (n=28)	Blended (n=28)	Total
Remember	.04 (.19)	.04 (.19)	.04 (.19)	1.25 (.84)	1.64 (1.28)	1.45 (1.09)
Understand	.00 (.00)	.25 (.59)	.13 (.43)	2.29 (2.11)	2.00 (1.27)	2.14 (1.73)
Analyze	.00 (.00)	.21 (.42)	.11 (.31)	2.89 (2.13)	2.68 (1.54)	2.79 (1.85)
Total	.01 (.11)	.17 (.43)	.09 (.33)	2.14 (1.90)	2.11 (1.42)	2.13 (1.67)

Neither task complexity nor the interaction between task complexity and vertical display had a significant effect on the number of clicks on vertical results. However, it is interesting to note that for more complex tasks, almost all vertical clicks were made on the blended interface. In fact, of all 15 vertical clicks (across all tasks, subjects, and interfaces), only one was made on the non-blended interface. These results (visualized in Figure 3) suggest that during more complex tasks, subjects tend to examine more vertical results. However, they are prone to do so only when the vertical results are blended. While we consider this trend worthy of further investigation, the overall number of vertical clicks was low. We revisit this point in Section 5.

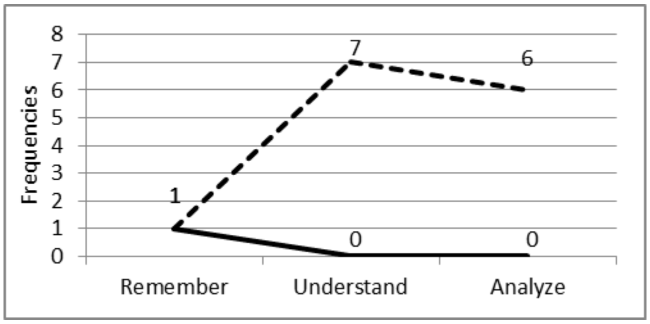


Figure 3. Comparison of vertical clicks on the blended (dashed line) and the non-blended interface (solid line)

4.3. Perceptions & Preferences

Our fourth research question explored users' evaluations of the interfaces, their perceptions of the verticals, and their display preferences.

After completing each set of tasks with one interface, subjects were asked to evaluate the system. Subjects' responses to the User Experience Questionnaire are shown in Table 8. There was little difference in subjects' ratings of the systems. For Perceived Usability, the mean ratings were identical. Paired samples t-tests were conducted and none of these differences were significant [Focused Attention: *t*(27)=.953, *p*=.349; Felt Involvement: *t*(27)=.135, *p*=.894; Perceived Usability: *t*(27)=.000, *p*=1.000; Endurability: *t*(27)=-.325, *p*=.748; and Search Effectiveness: *t*(27)=-.747, *p*=.462].

Table 8. Subjects' evaluations of each system (Mean, St. Dev.)

User Experience Sub-Scale	**System**	
	Non-blended	Blended
Focused Attention	2.96 (0.75)	2.83 (0.68)
Felt Involvement	3.50 (0.75)	3.48 (0.71)
Perceived Usability	3.69 (0.99)	3.69 (0.85)
Endurability	3.79 (0.77)	3.86 (0.72)
Search Effectiveness	3.54 (0.82)	3.71 (0.90)

After completing both sets of tasks (one set with each interface), subjects were asked about their perceptions of verticals and vertical results as well as their interface preferences.

As shown in Figure 2, subjects were first asked if they noticed any differences between interfaces. Twenty-three (82%) subjects claimed to notice differences. When asked what differences they noticed, the majority of subjects mentioned that the content of the results were more relevant for one system, others mentioned that the categories of search results were different, and others believed that the blended interface flowed better. A few subjects perceived blended vertical results as being advertisements.

Subjects were then asked whether they noticed the vertical tabs displayed on the top and left navigational bars. Twenty subjects (71%) said they noticed the verticals. Of these, the majority of subjects expected that clicking on a vertical tab would display a specific type of search result. Some said they expected the search engine to narrow the search results. One subject stated that he/she was concerned he/she would miss information by using the verticals. The 8 (29%) subjects that did not initially notice the verticals stated similar expectations, believing that the verticals would break down the search results by category.

Of the 20 subjects who noticed the verticals, only 7 (35%) said they used the verticals while searching. Of these, 3 thought the verticals were useful and 4 did not think they were useful either because they did not retrieve the right information or because they did not retrieve any results. People who did not use the verticals expressed that the search tasks did not require verticals; one subject said he/she was too preoccupied by the tasks to use verticals; another said that he/she did not usually use verticals.

Finally, at the end of the exit questionnaire, subjects were asked to express their preferences between interfaces in terms of information quality, ease of use, and overall experience (Figure 4). The preferences expressed by most subjects (n=22, 79%) were consistent across all three questions (information quality, ease of use, and overall). Only one subject favored different interfaces for different questions. Regarding information quality, 11 (39%) subjects preferred the non-blended interface, 12 (43%) preferred the blended interface, and the remaining five (18%) felt that both interfaces provided equal quality of information. For ease of use, 10 (36%) preferred the non-blended interface, 12 (43%) preferred the blended interface, and 6 (21%) stated no preference. Finally, in terms of overall preference, 11 (39%) preferred the non-blended interface, 13 (47%) preferred the blended interface, and 4 (14%) stated no preference. None of the chi-square tests were significant (information quality: $X^2(1) = 0.43$, $p = .84$; ease of use: $X^2(1) = 0.18$, $p = .67$; overall: $X^2(1) = 0.17$, $p = .68$).

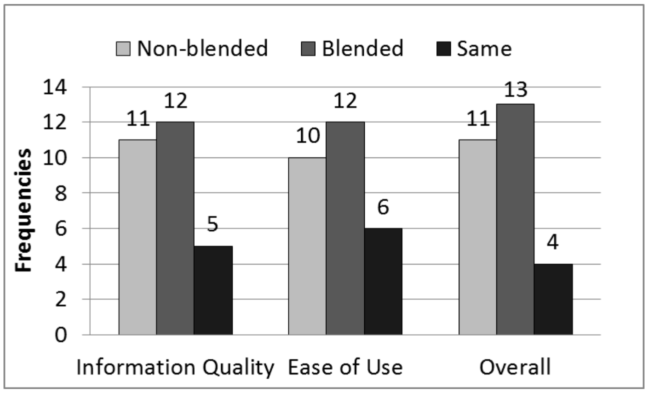


Figure 4. System Preferences

Given the clear division in system preferences, we reanalyzed subjects' evaluations of the systems (Table 8). Results showed that people who preferred the non-blended interface rated this interface higher than the blended on all the user experience sub-scales, while people who preferred the blended interface rated this interface higher than the non-blended interface (Table 9). These differences were especially pronounced for Perceived Usability and Endurability. We did not perform significance testing because of the small number of data points in each cell. However, these results suggest that the differences are not necessarily interface dependent, but rather person-dependent.

Table 9. Subjects' user experience ratings for each system according to their system preferences (Mean, St. Dev.)

User Experience Sub-Scale	People who preferred non-blended (n=11)		People who preferred blended (n=12)	
	Non-Blended	Blended	Non-Blended	Blended
Focused Attention	3.20 (.77)	2.83 (.70)	2.77 (.62)	2.83 (.74)
Felt Involvement	3.97 (.48)	3.06 (.78)	3.06 (.59)	3.75 (.75)
Perceived Usability	4.31 (.56)	2.83 (.73)	2.96 (.84)	4.08 (.46)
Endurability	4.32 (.30)	3.08 (.66)	3.25 (.70)	4.17 (.43)
Search Effectivenes s	3.8 (.83)	3.22 (.87)	3.32 (1.01)	4.02 (.73)

For each preference question, subjects were asked to justify their choices. The analysis of these responses showed a common set of rationales. Responses most often focused on the relevance of the information provided and the display of results. Because subjects often repeated their rationales across questions, responses are reported at the subject level to avoid misrepresenting the frequencies of certain rationales.

Subjects who preferred the non-blended system stated that it presented more relevant information (n=8), more credible information (n=3), and had a better layout (n=9). Subjects described the layout as "simple" and "intuitive" and stated that it helped them to more quickly identify relevant information (n=3) and was easier to use (n=4). Many subjects justified their preferences by commenting on what they did not like about the layout of the blended interface: "it spat everything out," "it was too visually overwhelming," and "it threw everything at you at once." Two subjects likened the verticals to commercial results stating that the non-blended interface had "fewer pulsing ads and screen garbage" and that the blended interface retrieved more "commercial results" from "less credible sources." These results illustrate how the display itself can influence judgments of information quality, especially in cases when the vertical results might not be relevant to the search query. Finally, one subject attributed her preference to experience, "I thought the [non-blended] was easier to navigate for me as a novice."

Subjects who preferred the blended system used similar rationales. Nine stated that the blended system provided more relevant information, although only one stated that it provided more credible information. With respect to layout, two subjects stated that the layout helped them to more quickly identify relevant information while one person stated that the system returned results faster. Comments regarding layout and ease of use were opposite of those expressed by subjects who preferred the non-blended system. Subjects who preferred the blended system thought the display was "visually pleasing" and "easier to use and more pleasant to look at." Subjects also liked that different types of results were presented and that they were "broken down into categories." Subjects commented that the display made "the subject matter easier to read and attain" and helped them be more efficient by "identifying different types of results." One subject described the vertical results as "illustrations," stating that "the illustrations made finding the information quicker and more entertaining." One subject justified his/her preference by stating that this display was the one which he/she was most familiar.

5. DISCUSSION

With respect to RQ1, more complex tasks required greater levels of search interaction: longer search sessions, more queries, more clicks on search results, and more webpages visited. Our results are consistent with results from Liu *et al* [13], who used three levels of objective task complexity, and consistent with Jansen *et al* [20], who used six levels of cognitive complexity (*remember*, *understand*, *apply*, *analyze*, *evaluate*, and *create*). Interestingly, Jansen *et al* [20] found that *evaluate* and *create* tasks required less search interaction than *analyze* tasks. However, with respect to the same cognitive complexity classes investigated in this work (*remember*, *understand,* and *analyze*), the trends are consistent.

When examining the effect of task complexity on the use of vertical results (RQ2), we observed a trend towards more clicks on vertical results for more complex tasks. However, the relationship was not significant. Our initial idea was to test five levels of task complexity (*remember*, *understand, analyze*, *evaluate*, and *create*), the hypothesis being that more complex tasks (and, in particular, the most complex: *evaluate* and *create*) would require increasing amounts and diversity of information to complete and would therefore require more use of vertical results. However, we discovered in our pilot tests that our target subjects could not complete 10 tasks in the time allotted (one task for each task-complexity/interface pair) and therefore decided to adjust the study design to include only tasks representing the first three levels of complexity (*remember, understand*, and *analyze*). While we did not find a significant main effect of task complexity on vertical usage, we found an interesting interaction between task complexity and vertical display (though, not significant, $p = 0.07$). Subjects clicked on more verticals when completing *understand* and *analyze* tasks with the blended interface. It is possible that we will find a significant effect of task complexity on vertical usage as well as an even stronger interaction effect between task complexity and vertical display when we examine the full range of task complexities, which we plan to do in future research.

Our findings regarding preferences between interfaces (RQ4) differ from Sushmita *et al.* [24], who found that subjects rated a blended interface more favorably than one where vertical results needed to be accessed by clicking on tabs. This difference might be the result of different study samples—Sushmita *et al.* [24] recruited subjects from the university and had a sample consisting of students, post-graduates and research staff. On the whole, these subjects were likely to be more technologically savvy than our subjects. Although Sushmita *et al.* [24] did not report the age of their subjects, they were likely younger than our subjects whose average age was 42 years. Moreover, the search tasks used in Sushmita *et al.* [24] implicated verticals, so this might also explain the difference.

Within our study, differences between subjects might also explain some of the variations in subjects' preferences and even in their use of verticals. For example, subjects with less search experience might prefer, and perform better with the non-blended interface because the results display is more parsimonious. Subjects with less search experience might find the blended display too visually overwhelming and distracting; this in turn, might impact their abilities to identify relevant information. While we measured Search Self-Efficacy (SSE), we did not initially design our experiment to test this variable. We conducted some preliminary analyses using this variable, but were not able to draw definitive conclusions because the search task domains were not distributed equally across the observed SSE scores and the SSE scores were slightly skewed. Future research should explore the potential relationship between search experience and vertical search display, along with other individual variables: people who are more visual thinkers might prefer the blended display or people who are novice users and/or have attention difficulties might prefer the non-blended display. If there are differences, then it might be useful for commercial search services to allow users to have more control over how verticals are displayed including the ability to turn-off verticals from a blended display.

6. CONCLUSION

We explored the relationship among task complexity, vertical display, and user interaction in aggregated search with research participants from our local community. Consistent with previous research, we found that more complex tasks were associated with greater interaction (i.e., longer sessions, more queries, more SERP clicks, and more pages visited). We also found that more complex tasks were associated with more clicks on vertical results. This is also consistent with previous research, which found that subjects used more sources during more complex tasks. Interestingly, however, users clicked on more vertical results during more complex tasks *only* when the vertical results were blended into the Web results. Finally, we found that subjects were divided in their preferences for vertical search displays. Differences between users (e.g., their search experience) may be an important factor in whether the blending of vertical results enhances or hinders the user experience. Future work will investigate tasks of additional complexity, a larger variety of presentation methods for verticals, and the effects of individual differences on display preferences.

This line of research has important implications for aggregated search systems. It may be worthwhile to consider *predicted* task complexity (predicted using session-level interaction signals) in vertical selection and presentation decisions. Furthermore, with respect to user preferences, personalization in aggregated search may be an interesting avenue to explore.

7. REFERENCES

[1] Anderson, L. W. and Krathwohl, D. A. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

[2] Arguello, J., Diaz, F., and Callan, J. (2011). Learning to aggregate vertical results into web search results. In *Proceedings of CIKM*. ACM, 201–210.

[3] Arguello, J., Diaz, F., Callan, J. and Carterette, B. (2011). A methodology for evaluating aggregated search results. In *Proceedings of ECIR*. Springer-Verlag, 141–152.

[4] Arguello, J., Diaz, F., Callan, J. and Crespo, J.-F. (2009). Sources of evidence for vertical selection. In *Proceedings of SIGIR*. ACM, 315–322.

[5] Arguello, J., Diaz, F., and Paiement, J.-F. (2010). Vertical selection in the presence of unlabeled verticals. In *Proceedings of SIGIR.* ACM, 691-698.

[6] Bell, D. & Ruthven, I. (2004). Searcher's assessments of task complexity for Web searching. In *Proceedings of ECIR*. Springer-Verlag, 57–71.

[7] Byström, K. and Hansen, P. (2005). Conceptual framework for tasks in information studies. *Journal of the American Society for Information Science and Technology, 56*(10), 1050–1061.

[8] Byström, K. and Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing and Management, 31*(2), 191–213.

[9] Campbell, D. J. (1988). Task complexity: A review and analysis. *Academy of Management Review, 13*(1), 40–52.

[10] Debowski, S., Wood, R., and Bandura, A. (2001). The impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic enquiry. *Journal of Applied Psychology, 86*(6), 1129–1141.

[11] Diaz, F. (2009). Integration of news content into web results. In *Proceedings of WSDM*. ACM, 182–191.

[12] Hackman, J. R. (1969). Toward understanding the role of task in behavioral research. *Acta Psychologica,* 31, 162–187.

[13] Jansen, B. J., Booth, D., & Smith, D. (2009). Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management, 45*, 643–663.

[14] Kellar, M., Watters, C., and Shepherd, M. (2007). A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for Information Science & Technology, 58*(7), 999–1018.

[15] Kim, J. (2006). Task difficulty as a predictor and indicator of web searching interaction. In *Proceedings of CHI*. ACM, 959–964.

[16] Li,Y. (2009). Exploring the relationships between work task and search task in information search. *Journal of the American Society for Information Science & Technology, 60*(2), 275–291.

[17] Li, Y. and Belkin, N. J. (2008). A faceted approach to conceptualizing tasks in information seeking. *Information Processing & Management, 44*, 1822–1837.

[18] Li, Y. and Belkin, N. J. (2010). An exploration of the relationship between work task and interactive information search behavior. *Journal of the American Society for Information Science & Technology, 61*(9), 1771–1789.

[19] Li, X., Wang, Y.-Y., and Acero, A. (2008). Learning query intent from regularized click graphs. In *Proceedings of SIGIR*. ACM, 339–346.

[20] Liu, J., Cole, M., Liu, C., Bierig, R. Gwizdka, J., Belkin, N.J., Zhang, J., and Zhang, X. (2010). Search behaviors in different task types. In *Proceedings of JCDL*. ACM, 69–78.

[21] O'Brien, H. L. (2010). The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping. *Interacting with Computers, 22*, 344–352.

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

[23] Ponnuswami, A., Pattabiraman, K., Wu, Q., Gilad-Bachrach, R., and Kanungo, T. (2011). On composition of a federated web search result page: Using online users to provide pairwise preference for heterogeneous verticals. In *Proceedings of WSDM*. ACM, 715–724.

[24] Sushmita, S., Joho, H., and Lalmas, M. (2009). A task-based evaluation of an aggregated search interface. In *Proceedings SPIRE*. Springer–Verlag, 322–333.

[25] Sushmita, S., Joho, H., Lalmas, M., and Villa, R. (2010). Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of CIKM*. ACM, 519–528.

[26] Toms, E. G. (2011). Task-based information searching and retrieval. In I. Ruthven & D. Kelly (Eds.) *Interactive Information Seeking, Behaviour and Retrieval*. Facet Publishing, 43–59.

[27] Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science & Technology, 37*, 413–464.

[28] Zhu, D. and Carterette, B. (2010). An analysis of assessor behavior in crowdsourced preference judgements. In *Proceedings of SIGIR Workshop on Crowdsourcing for Search Evaluation*. ACM, 21–26.