The Effects of Vertical Rank and Border on Aggregated Search Coherence and Search Behavior

Jaime Arguello and Robert Capra

School of Information & Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, USA
[jarguello, rcapra]@unc.edu

ABSTRACT

Aggregated search is the task of blending results from different search services, or *verticals*, into a set of web search results. Aggregated search coherence is the extent to which results from different sources focus on similar senses of an ambiguous or underspecified query. Prior work investigated the "spill-over" effect between a set of blended vertical results and the web results. These studies found that users are more likely to interact with the web results when the vertical results are more consistent with the user's intended query-sense. We extend this prior work by investigating three new research questions: (1) Does the spill-over effect generalize across different verticals? (2) Does the vertical rank moderate the level of spill-over? and (3) Does the presence of a border around the vertical results moderate the level of spill-over? We investigate four different verticals (images, news, shopping, and video) and measure spill-over using interaction measures associated with varying levels of engagement with the web results (bookmarks, clicks, scrolls, and mouseovers). Results from a large-scale crowdsourced study suggest that: (1) The spill-over effect generalizes across verticals, but is stronger for some verticals than others, (2) Vertical rank has a stronger moderating effect for verticals with a mid-level of spill-over, and (3) Including a border around the vertical results has a subtle moderating effect for those verticals with a low level of spill-over.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Storage and Retrieval

General Terms

Performance, Experimentation, Human Factors

Keywords

Aggregated search; user study; search behavior; evaluation

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2661930.

1. INTRODUCTION

In addition to web search, commercial search portals such as Google, Bing and Yahoo! provide access to a wide range of specialized search services or *verticals*. Common verticals include search engines for a specific type of media (e.g., images, videos, books) or a specific type of search task (e.g., search for news, local businesses, on-line products). The goal of *aggregated search* is to combine results from these different systems in a single presentation. From a system perspective, aggregated search is a two-part task: (1) predicting which verticals to present in response to a query (*vertical selection* [4, 5, 11, 19]) and (2) predicting where to present those verticals selected (*vertical presentation* [3, 23, 24]). Typically, a vertical is presented by blending a few of its top results somewhere in the first page of web results. The most confidently relevant verticals are blended higher.

In this work, we study a phenomenon called *aggregated search coherence* [1]. Given an ambiguous query (e.g., "saturn"), a common strategy for a search engine is to diversify its results (e.g., to return results about Saturn the planet, the car, and the Roman god). Aggregated search coherence is the extent to which results from different sources focus on similar query-senses. Suppose that a user issues the query "saturn" and the system decides to blend image vertical results into the web results. If the web results cover multiple senses, but the images focus exclusively on the planet, then the aggregated results have a *low* level of coherence. Conversely, if both sets of results cover the same query-senses, then the aggregated results have a *high* level of coherence.

We investigate the effect of aggregated search coherence on search behavior. Specifically, we consider whether the query-senses represented in a set of vertical results blended on the search results page (SERP) can affect user interaction with the web results. Prior work by Arguello and Capra studied the "spill-over" effect between images and web results [1]. Results showed that users are more likely to interact with the web results when the blended images are more consistent with the intended query-sense. That is, a user looking for web results about "saturn" the planet is more likely to interact with the web results if the images contain pictures of the planet versus the car. In a follow-up study, Arguello and Capra investigated whether the spill-over effect generalizes across verticals [2]. Results showed a strong spill-over for the images and shopping verticals, a moderate spill-over for the video vertical, and no spill-over for the news vertical. We build upon this prior work and explore three new research questions.

Measuring spill-over requires detecting whether the query-senses in the vertical results affect user engagement with the web results. Arguello and Capra [1, 2] operationalized user engagement with the web results in terms of clicks and bookmarks.[1] While these are strong signals of user engagement, their absence does not necessarily indicate zero engagement. Thus, our first research question (RQ1) re-visits whether the spill-over effect generalizes across verticals. We investigate four different verticals (images, news, shopping, and video) and, to measure engagement with the web results, we complement clicks and bookmarks with additional signals derived from browsing behavior (scrolls and mouseovers). We treat scrolls and mouseovers as evidence that the user scanned the web results and study whether the query-senses in the vertical results influence this behavior.

Our second and third research questions investigate two factors that may moderate the level of spill-over from the vertical to the web results. An important objective in aggregated search is to make the most confidently relevant verticals more salient. One approach is to blend them higher in the web results. Our second research question (RQ2) investigates whether the vertical rank moderates the level of spill-over. Prior eye-tracking studies found that users spend more time scanning results that are higher on the SERP [7, 9, 12, 15, 22]. Some of these studies also show a *trust bias* in favor of higher-ranked results [15, 22]. Within aggregated search, higher-ranked verticals may cause more spill-over for two reasons. First, users may be more likely to notice the vertical and thus the query-senses in the vertical results. Second, users may assume that the system is more confident about the majority query-sense in the vertical results and that the web results are also skewed towards this sense.

In most commercial systems, vertical results are blended into the web results without a strong visual cue to separate them from the web results. In contrast, advertisements are often displayed using a different colored background or are enclosed in a border. Such visual cues subtly communicate to a user that the ads are separate from the web results and should be treated differently. Our third research question (RQ3) investigates whether the presence of a thin (2-pixel) gray border around the vertical results moderates the level of spill-over from the vertical to the web results. A border may reduce the level of spill-over for two reasons. First, a border may signal to a user that the vertical results are a different *type* of result. If the user wants web results instead, a border may influence them to skip-over the vertical and not even recognize the query-senses in the vertical results. Second, even if a user notices the query-senses in the vertical results, a border may convey that the vertical results come from a different source than the web results, which may influence the user to judge the web results independently from the vertical results. We investigate these three research questions in a large-scale crowdsourced user study.

2. RELATED WORK

Current methods for vertical selection and presentation do not *explicitly* favor coherent results. This is true in terms of the evidence used by existing algorithms to make predictions and the evaluation methods used to measure performance.

Algorithms for vertical selection and presentation use machine learning to combine a wide range of features. Prior work investigated features generated from the query string [4, 11, 23, 24], from the vertical results [3, 4, 10, 11], from the vertical query-log [3, 4, 10, 11], and from historic click-through rates on the vertical results [23, 24]. None of the features reported in the literature consider the relationship between the vertical results and those from other components on the SERP.

Evaluation methods for vertical selection and presentation also ignore the effects of the vertical results on other components on the SERP. The goal of vertical selection is to make binary predictions for a set of candidate verticals. Thus, selection algorithms are evaluated using vertical relevance judgements and metrics such as accuracy [4, 5, 11] or precision and recall for each vertical [19]. A limitation of these metrics is that all false positive predictions are treated equally. In this study, we show that depending on the vertical results, displaying a non-relevant vertical can also affect user interaction with other components on the SERP.

Evaluation methods for vertical presentation fall under three categories: on-line, test-collection, and whole-page evaluation methods. On-line methods are used to evaluate systems in a live environment using implicit feedback [10, 24]. If a vertical is presented, a vertical *click* signals a true positive prediction and a *skip* signals a false positive prediction. These methods suffer from the same limitation mentioned above—a vertical skip may deserve special treatment if it also affects user interaction with other components. Test-collection methods follow a Cranfield-style evaluation [8]. A test-collection includes a set of queries, cached results from different sources, and relevance judgements. Zhou *et al.* [30] proposed a utility-based evaluation metric that considers three distinguishing properties between verticals: (1) the vertical's relevance to the task, (2) the likelihood of a user noticing the vertical results, and (3) the expected effort required to assess their relevance. Our work suggests a fourth aspect to consider: the expected spill-over from the vertical results to other components on the SERP. Bailey *et al.* [6] proposed a whole-page evaluation method called SASI, which elicits human judgements on the whole SERP. While cross-component coherence is mentioned as an important aspect of whole-page quality, its effect on user behavior and satisfaction was not investigated [6].

Incoherent results occur when the query is ambiguous and the different aggregated components focus on different query-senses. A natural question is: How often does this happen? Sanderson [25] conducted an analysis of a large commercial query-log and found that roughly 4% of all unique queries and 16% of all unique head queries (the most frequent) had an exact match with an ambiguous entity in Wikipedia and WordNet. Given an ambiguous query, incoherent results are more likely when the different back-end collections are skewed towards different senses. The analysis by Santos *et al.* [26] suggests that this is often the case. Santos *et al.* considered the different senses for a set of ambiguous entities and compared their frequencies in query-logs from a commercial web search engine and three verticals. Results showed that different sources were skewed towards different senses. For example, the shopping vertical had more queries about "amazon" the company, while the images vertical had more queries about the rainforest.

Our first research question (RQ1) investigates whether the spill-over effect generalizes across verticals. Arguello and Capra [1, 2] considered the same research question, but

[1]Clicking a web result or marking it as relevant.

measured spill-over only in terms of bookmarks and clicks. We extend this work by considering additional interaction measures that also indicate engagement with the web results (scrolls and mouseovers). Outside of aggregated search, prior work considered the spill-over from advertisements on the SERP to the web results. Kalyanaraman and Ivory [16] found that relevant ads improve users' attitudes towards the web results, and Buscher *et al.* [7] found that relevant ads increase the visual attention paid to the web results.

Our second research question (RQ2) investigates whether the vertical rank moderates the level of spill-over. Eye-tracking studies show that users tend to focus their visual attention on results that are ranked higher on the SERP [7, 9, 12, 15, 22]. Prior studies also found a *trust bias* in favor of higher ranked results [15, 22]. Users had a click-bias in favor of the top result, even in cases where the they scanned the top-two and the second was more relevant. Within aggregated search, Sushmita *et al.* [28] found a click-bias in favor of higher-ranked verticals.

Our third research question (RQ3) investigates whether including a border around the vertical results moderates the level of spill-over. Designing effective information displays such as SERPs requires understanding how people perceive groups of objects [29]. Early work in psychology developed the Gestalt principles of pattern recognition [18]. For example, the Gestalt principles of *similarity* and *proximity* state that items in a display that are similar and/or closer together are perceived as a group. Palmer [21] later proposed the principle of *common region*, which states that items displayed in a common region (such as within a border) are perceptually grouped. Commercial search portals use the principle of common region to distinguish certain elements on the SERP (e.g., advertisements) from the web results. However, the use of common region is not typically applied to verticals that are blended into web results. Additionally, while search engine companies test the effects of subtle layout changes on search behavior [13], the effect of common region on the level of spill-over between components on the SERP has not been reported in prior work.

3. METHODS AND MATERIALS

We investigated our three research questions in a large-scale crowdsourced user study. Next, we describe the experimental protocol (Section 3.1), our experimental variables (Section 3.2), the search tasks used (Section 3.3), and some implementation details (Section 3.4).

3.1 User Study Overview

The experimental protocol is shown in Figure 1. Participants were given a search task and asked to use a live search engine to find a webpage containing the requested information. Search tasks had the form "Find information about `<entity>`", for example "Find biographical information about the British politician Michael Moore." The search engine was implemented using the Bing Web Search API.

Our goal was to study search behavior under the following scenario: First, a user has a search task in mind (e.g., "Find information about the plot of the movie Salt.") and issues to the system an ambiguous query (e.g., "salt"). Then, in response to this query, the system decides to show results from a particular vertical in addition to the web results and returns vertical results with a particular query-sense distribution (e.g., all images about "salt" the mineral). Finally,

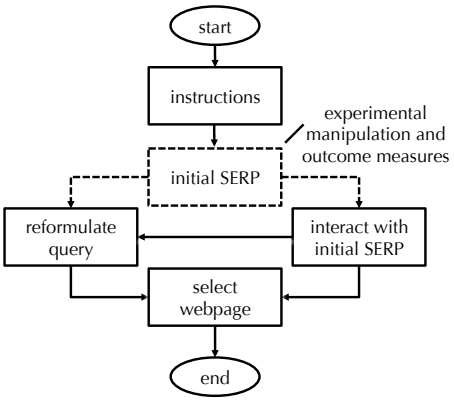


Figure 1: Experimental protocol.

based on these results, the user must decide whether to interact with the web results or reformulate the query. Our basic research questions are whether the query-senses in the vertical results influence the user's decision to interact with the web results and whether the vertical rank and the presence of a border around the vertical results have a moderating effect. In order to do a controlled study of the scenario described above, participants were told that "to help get you started with the search task, you will be provided with an initial query and a set of results." We refer to this starting point SERP as the *initial SERP*. This initial SERP is where the experimental manipulation took place.

The study proceeded as follows. After reading a set of instructions, participants were routed to the initial SERP. The initial SERP included the search task description, an initial query, and an initial set of results, supposedly returned by the system in response to the initial query. As described in more detail below, the initial query was purposely ambiguous and the initial results included web results and a set of blended results from one of four verticals (images, news, shopping, and video). The web results corresponded to the top-10 results returned by the Bing Web Search API and the vertical results were experimentally manipulated as described in Section 3.2. From the initial SERP, participants were instructed to search naturally by examining the results provided in the initial SERP or by issuing their own queries. Participant queries returned results using the Bing Web Search API and did not include vertical results. Clicking on a search result opened the landing page inside an HTML frame with a button above the frame labeled: "Click here if this webpage contains the requested information." Clicking this button ended the search task. At an point, participants were allowed to use the browser back button to return to the SERP and continue searching. The purpose of the study was disguised by telling participants that we were testing a new search engine.

Our goal was to study whether the vertical results influence user interaction with the web results. Thus, all our experimental outcome measures were derived from user interactions with the web results on the initial SERP. We focused on four binary outcome measures: (1) Did the participant select a web result on the initial SERP as containing the requested information? (selected web), (2) Did the participant click on a web result on the initial SERP? (clicked web), (3) Did the participant scroll down to see results be-

low the fold? (scroll), and (4) Did the participant mouseover on a web result on the initial SERP? (mouseover). These four outcome measures indicate varying levels of engagement with the web results on the initial SERP and extend prior work that considered only clicks and bookmarks [1, 2].

3.2 Experimental Variables

We manipulated four experimental variables: vertical, vertical query-sense, vertical rank, and vertical border. These variables controlled which vertical was blended into the initial SERP and how it was displayed. Figure 2 shows four example initial SERPs.

The *vertical* variable determined which vertical was blended into the initial SERP: images, news, shopping, or video. The images, news, and video verticals were implemented using search APIs provided by Bing, and the shopping vertical was implemented using the product search API provided by eBay. For images, video, and shopping, we blended five vertical results horizontally, and for news, we blended three vertical results vertically. The blended vertical results were designed to look similar to commercial systems: image results were displayed using thumbnails; news results were displayed using the title, summary snippet, news source, publication age, and a reduced-size rendering of an image pulled from the article; shopping results were displayed using the product name, price, condition (new, used), and a picture of the product; and video results were displayed using the title, duration, and a keyframe of the video.

The *vertical query-sense* variable manipulated the query-senses represented in the vertical results and had three different values: on-target, off-target, and mixed. On target results were all consistent with the search task description, off-target results were all on a different sense, and mixed results had a combination of on-target and off-target results. For example, for the search task "Find information about the plot of the movie Salt.", on-target results were all about "salt" the movie, off-target results were all about "salt" the mineral, and mixed results had a combination of both. For images, shopping, and video, mixed results had two on-target and three off-target results. For the news vertical, mixed results had one on-target and two off-target results.

The *vertical rank* variable manipulated the position of the vertical and had two values: rank one and rank four. For rank one, the vertical was positioned directly above the first web result, and for rank four, the vertical was blended between the third and fourth web results.

The *vertical border* variable manipulated the presence of a thin (2-pixel) gray border around the vertical results and had two values: border and no-border.

3.3 Search Tasks

Each vertical was associated with its own unique set of 60 search tasks and each search task was associated with four components: (1) the search task description, (2) the initial query, (3) the on-target sense, and (4) the off-target sense. We used a subset of the search tasks used in Arguello and Capra [2] and the same cached web and vertical results used to construct the initial SERPs. We provide an overview of how the search tasks were constructed and refer the reader to Arguello and Capra [2] for more details.

The tasks were created as follows. First, a large set of candidate initial queries was constructed by identifying all entities with a Wikipedia disambiguation page. Second, in order to use initial queries that a user might realistically enter as a query, all entities not appearing as a query in the AOL query-log were filtered. The third step was to identify initial queries with a strong orientation towards one of the four verticals. To this end, the remaining entities were issued to Bing and four (possibly overlapping) subsets of entities were constructed based on whether the entity triggered the images, news, shopping, or video vertical in the Bing results. The next step was to identify entities that returned multiple senses from its corresponding vertical. To accomplish this, the authors manually inspected the results from each vertical and selected 75 entities per vertical that returned multiple senses in the top-50 results. One of these senses was assigned to be the on-target sense and another was assigned to be the off-target sense. Finally, the search task description was constructed to be consistent with the on-target sense. Because the goal was to study the spill-over from the vertical to the web results, the search tasks were designed to favor web results. We randomly sampled a subset of 60/75 search tasks per vertical to use in our study. Table 1 shows a few example tasks.

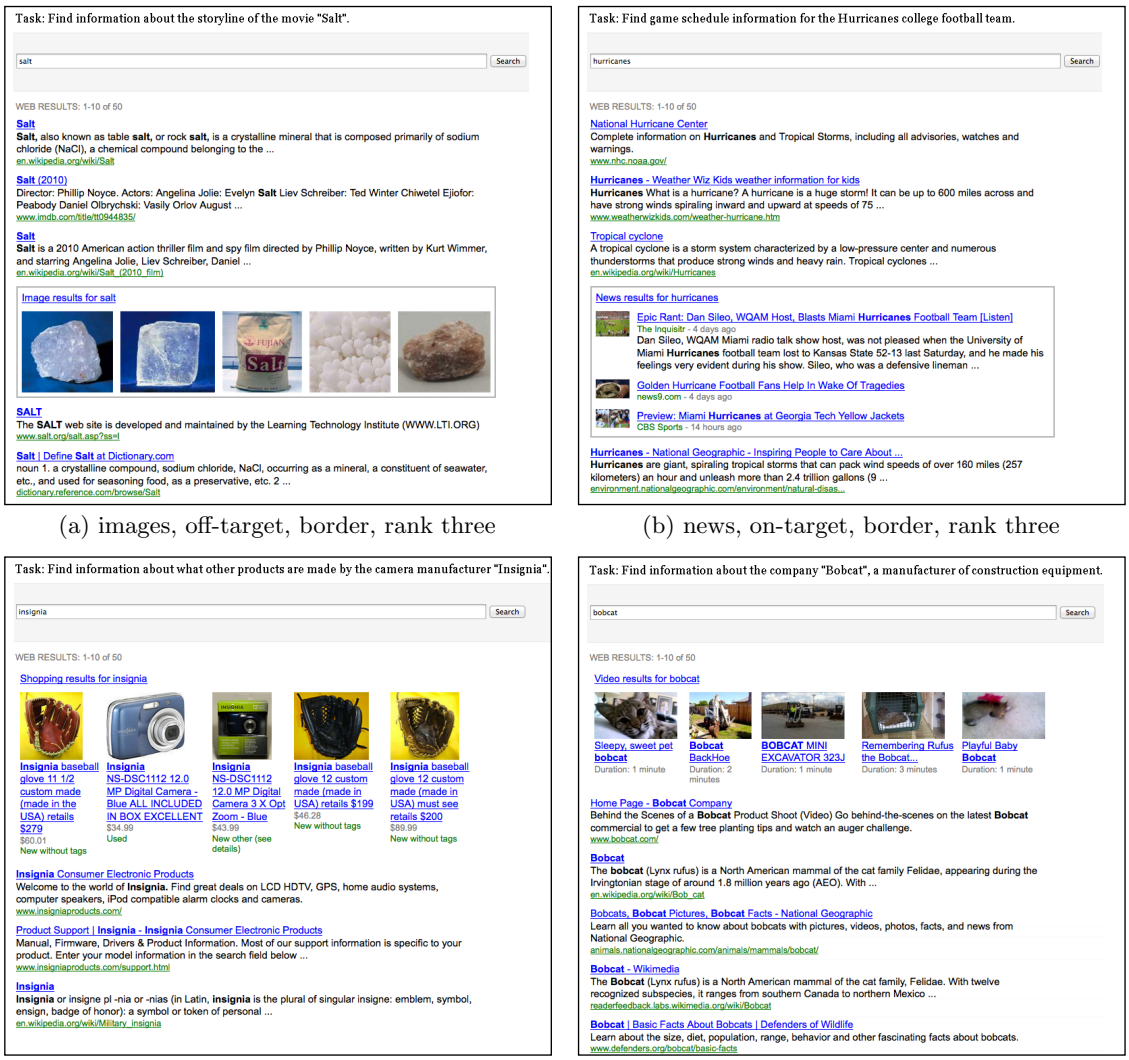
3.4 User Study Details

The study was run as a remote study using Amazon's Mechanical Turk (MTurk). Each MTurk Human Intelligence Task (HIT) was associated with a single search task. Each vertical had its own unique set of 60 search tasks, and within each vertical, the study design was fully crossed. This resulted in $2,880$ experimental conditions (4 verticals $\times$ 60 search tasks per vertical $\times$ 3 vertical query-senses $\times$ 2 vertical ranks $\times$ 2 vertical border conditions = 2,880). Finally, for each experimental condition, we collected data from 10 participants, for a total of 28,880 HITs. Each HIT was priced at $0.10 USD.

Our HITs were implemented as *external* HITs, meaning that everything besides recruitment and compensation was managed by our own server. Hosting our HITs externally allowed us to capture all the necessary user interactions with the initial SERP, control the assignment of workers to experimental conditions, and do quality control dynamically. All our outcome measures were derived from user interactions with the initial SERP (Section 3.1). Clicks were captured at the server-side using URL re-directs, and scrolls and mouseovers were captured at the client-side using Javascript and transmitted to our server using AJAX.

MTurk workers were assigned to experimental conditions randomly, except for three constraints. First, participants were not allowed to see the same search task more than once. Second, vertical border was a between-subjects variable. We were concerned that varying the presence of a border as a within-subjects variable might water-down its effect, limiting our ability to detect differences in the two conditions. Thus, participants were assigned to a single vertical border condition. Finally, in order to collect data from a wide range of participants, workers not allowed to complete more than 100 of our HITs (less than 0.05% of those available). In total, we collected data from 898 participants.

MTurk studies require quality control, and we addressed this in three ways. First, we restricted our HITs to workers with a 95% acceptance rate or greater. Second, to help ensure English language proficiency, we restricited our HITs to workers in the U.S. Finally, using an external HIT design allowed us to do quality control dynamically. Prior to the



(a) images, off-target, border, rank three

(b) news, on-target, border, rank three

(c) shopping, mixed, no border, rank one

(d) video, mixed, no border, rank one

Figure 2: Example initial SERPs for different experimental conditions (cropped).

experiment, one of the authors judged the relevance of every web result on an initial SERP. Workers who selected three non-relevant results from an initial SERP as containing the requested information were not allowed to do more HITs.

4. RESULTS

In the following sections, we examine the main effect of vertical query-sense (Section 4.1), the main effect of vertical rank (Section 4.2), the main effect of vertical border (Section 4.3), the interaction effect between vertical rank and query-sense (Section 4.4), and the interaction effect between vertical border and query-sense (Section 4.5).

4.1 Vertical Query-sense

In this section, we examine the level of spill-over across all four verticals. User interaction with the web results from the initial SERP was operationalized using the four measures previously described: selected web, clicked web, scroll, and mouseover. For each interaction measure, the presence of a spill-over effect is indicated by an increase in the probability of the interaction being true when the vertical results are more on-target (i.e., on-target vs. mixed vs. off-target).

Figures 3(a)-3(d) show the results for each vertical and each query-sense condition: off-target (off), mixed (mix), and on-target (on). The percentages shown indicate the percentage of trials for which the interaction measure was true (i.e., the user selected, clicked, scrolled, and mouseovered). Recall that each vertical had its own unique set of 60 tasks and that each experimental condition was completed by 10 participants. Thus, each percentage was computed across 2,400 trials (60 tasks $\times$ 2 rank conditions $\times$ 2 border conditions $\times$ 10 participants per condition = 2,400).

The images vertical had the strongest level of spill-over. Chi-squared tests showed significant main effects of vertical query-sense on all four interaction measures: selected web ($\chi^2(2) = 20.661$, $p < .001$), clicked web ($\chi^2(2) = 35.043$, $p < .001$), scroll ($\chi^2(2) = 19.907$, $p < .001$), and mouseover ($\chi^2(2) = 6.612$, $p < .05$). Post-hoc comparisons found significant differences for selected web (on/off, mix/off), clicked web (on/off, mix/off), scroll (on/mix, on/off), and mouseover

Table 1: Example search tasks.

vertical	initial query	on-target sense	off-target sense	task description
images	washington	state	historical figure	Find information about tourist attractions in Washington State.
images	proton	car	particle	Find information about the Proton automobile company.
images	wega	coffee maker	sony tv	Find information about the coffee maker company 'Wega'.
news	sierra nevada	beer maker	mountain	Find information about where Sierra Nevada beer is made.
news	big bang theory	scientific theory	tv show	Find information about the history of the Big Bang Theory of the universe.
news	anthrax	disease	band	Find information about different modes of infection for anthrax.
shopping	twister	game	movie	Find historical information about the popular game 'Twister'.
shopping	bladerunner	movie	skates	Find information about the book that inspired the movie Bladerunner.
shopping	the eagles	band	football team	Find biographical information about the band 'The Eagles'.
video	blody mary	song	cocktail	Find the lyrics to the song 'Bloody Mary' by Lady Gaga.
video	blue orchid	song	flower	Find the lyrics to the song 'Blue Orchid' by the White Stripes.
video	dark angel	band	tv show	Find biographical information about the band 'Dark Angel'.

(on/off).[2] The largest spill-over for images was for clicked web. Participants were 23% more likely to click on a web result from the initial SERP when the images were on-target versus off-target (an increase from 34.76% to 42.83%).

The shopping vertical had the second strongest level of spill-over. Again, chi-squared tests showed significant main effects of vertical query-sense on all interaction measures: selected web ($\chi^2(2) = 15.850$, $p < .001$), clicked web ($\chi^2(2) = 20.284$, $p < .001$), scroll ($\chi^2(2) = 11.704$, $p < .05$), and mouseover ($\chi^2(2) = 10.348$, $p < .05$). Post-hoc comparisons found significant differences for selected web (on/off, mix/off), clicked web (on/off, on/mix, mix/off), scroll (on/off, on/mix), and mouseover (on/off, mix/off). Similar to images, the strongest spill-over for shopping was for clicked web. Participants were 19% more likely to click on an initial SERP web result when the shopping results were on-target versus off-target (an increase from 32.40% to 38.62%).

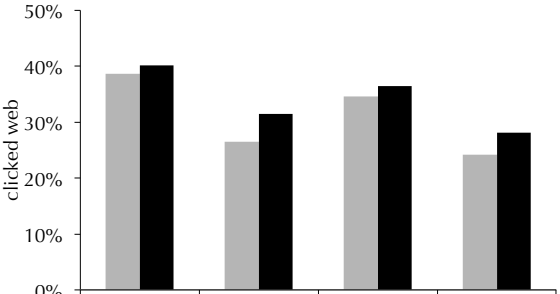
News and video had a similar trend. User interaction with the web results was greater when the news and video results were more on-target. However, the trend was less pronounced. For news, the level of spill-over was only significant for scroll ($\chi^2(2) = 8.052$, $p < .05$), with post-hoc comparisons showing significant differences between on/off and mix/off. For video, the level of spill-over was only significant for mouseover ($\chi^2(2) = 8.052$, $p < .05$), with post-hoc comparisons showing significant differences between on/off and mix/off.

4.2 Vertical Rank

In this section, we examine the main effect of vertical rank on user interaction with the web results. Due to space limitations, we focus this and all the following analyses on the clicked web outcome measure. Results are presented in Figure 4. Each percentage was computed across 3,600 trials (60 tasks × 3 query-sense conditions × 2 border conditions × 10 participants per condition = 3,600).

For all verticals, user interaction with the web results was greater when the vertical was presented at rank four versus rank one. However, while the trend was the same for all verticals, it was only significant for news ($\chi^2(1) = 21.632$, $p < .001$) and video ($\chi^2(1) = 14.558$, $p < .001$). The main effect of rank was not significant for images ($\chi^2(1) = 1.799$, $p = .180$) and shopping ($\chi^2(1) = 2.700$, $p = .100$).

[2]All post-hoc comparisons used the modified Bonferroni correction outlined in Keppel [17], p. 170.



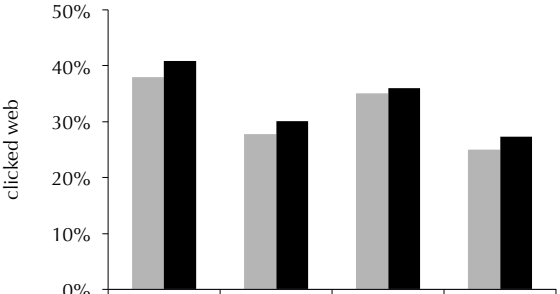
	images	news	shopping	video
rank 1	38.65%	26.47%	34.58%	24.17%
rank 4	40.19%	31.44%	36.44%	28.12%

Figure 4: Main effect of vertical rank

4.3 Vertical Border

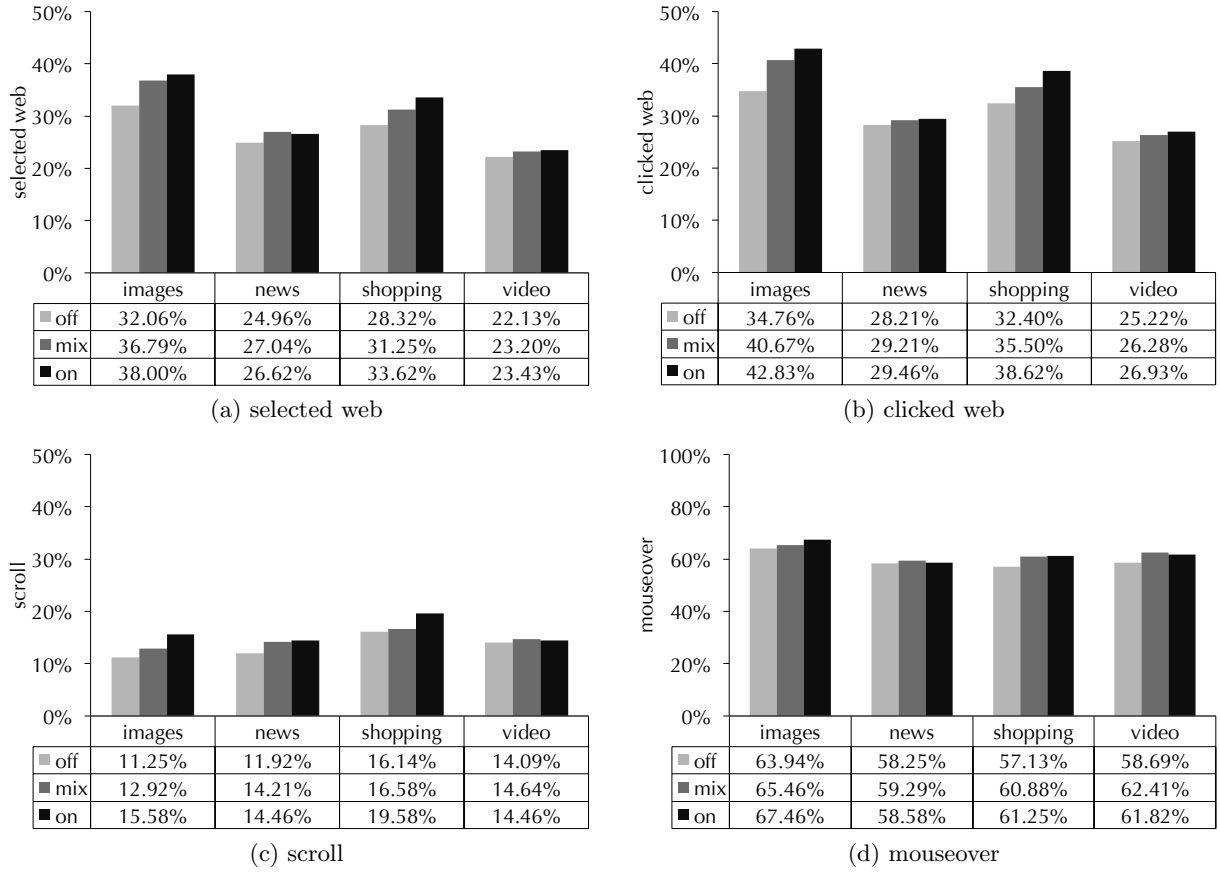
In this section, we examine the main effect of vertical border on user interaction with the web results. Results are presented in Figure 5. As before, each percentage was computed across 3,600 trials (60 tasks × 3 query-sense conditions × 2 rank conditions × 10 participants per condition = 3,600).

For all verticals, user interaction with the web results was greater in the presence of a border around the vertical results. The trend was the same for all verticals, and was significant for images ($\chi^2(1) = 6.005$, $p < .05$), news ($\chi^2(1) = 4.878$, $p < .05$), and video ($\chi^2(1) = 5.35$, $p < .05$). The main effect of border was not significant for shopping ($\chi^2(1) = 0.672$, $p = 0.412$).



	images	news	shopping	video
no border	38.01%	27.78%	35.05%	24.94%
border	40.83%	30.14%	35.97%	27.34%

Figure 5: Main effect of vertical border



(a) selected web (b) clicked web

(c) scroll (d) mouseover

Figure 3: Main effect of vertical query-sense

4.4 Vertical Rank and Query-sense

The analysis in Section 4.2 shows that user interaction with the web results was greater when the vertical was presented at rank four versus rank one. This general trend suggests that users engaged with the vertical results *more* when they were presented at rank one. In this section, we examine whether vertical rank moderated the level of spill-over from the vertical to the web results. In other words, we investigate if there was an interaction effect of vertical rank and query-sense. Results are presented separately for each vertical in Figures 6(a)-6(d). Each percentage was computed across 1,200 trials (60 tasks $\times$ 2 border conditions $\times$ 10 participants per condition = 1,200).

The level of spill-over is indicated by the magnitude of the upward slope from off-target to mixed and from mixed to on-target vertical results. The moderating effect of vertical rank on the level of spill-over can seen by comparing the upward slopes for rank four (solid line) and rank one (dashed line).

In general, the level of spill-over was greater when the vertical was presented at rank one versus rank four. For each vertical, we used a logistic regression to test the interaction between vertical query-sense and vertical rank. Both variables were treated as categorical. We used the off-target condition as the baseline for vertical query-sense and rank four as the baseline for vertical rank. Thus, each logistic regression included a total of five indicator variables: two for vertical query-sense, one for vertical rank, and two for their interaction. The only significant interaction was found for the shopping vertical. The difference on clicked web from on- versus off-target results was only significant at rank one (Wald's $\chi^2(1) = 4.166$, $p < .05$).

4.5 Vertical Border and Query-sense

The analysis in Section 4.3 found that user interaction with the web results was greater when there was a border around the vertical results. This general trend suggests that users noticed (and potentially processed) the vertical results *less* when they were enclosed in a border. In this section, we examine whether including a border moderated the level of spill-over from the vertical to the web results. Results are presented for each vertical in Figures 7(a)-7(d). Each percentage was computed across 1,200 trials (60 tasks $\times$ 2 rank conditions $\times$ 10 participants per condition = 1,200).

For all verticals except images, the level of spill-over (i.e., the magnitude of the upward slope from off-target to mixed and from mixed to on-target vertical results) was greater in the absence of a border. Similar to the previous section, for each vertical, we used a logistic regression to test the interaction between vertical query-sense and vertical border. We used the off-target condition as the baseline for vertical query-sense and the no border condition as the baseline for vertical border. Thus, each logistic regression included a total of five indicator variables: two for vertical query-sense, one for vertical border, and two for their interaction. None of the interaction variables were found to be significant.

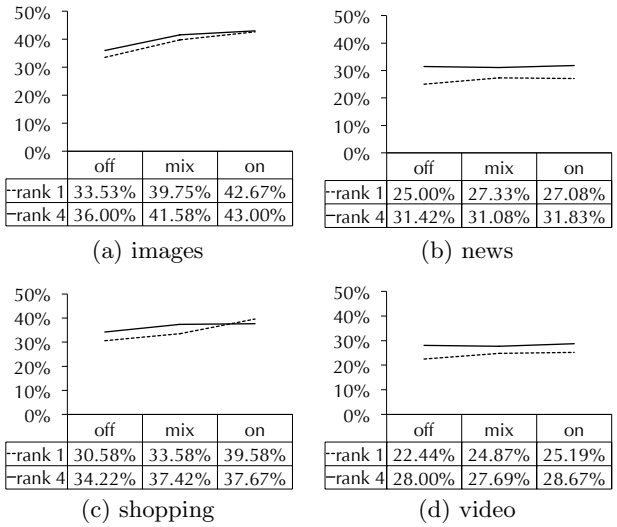


Figure 6: Interaction of rank and sense on clicked web.

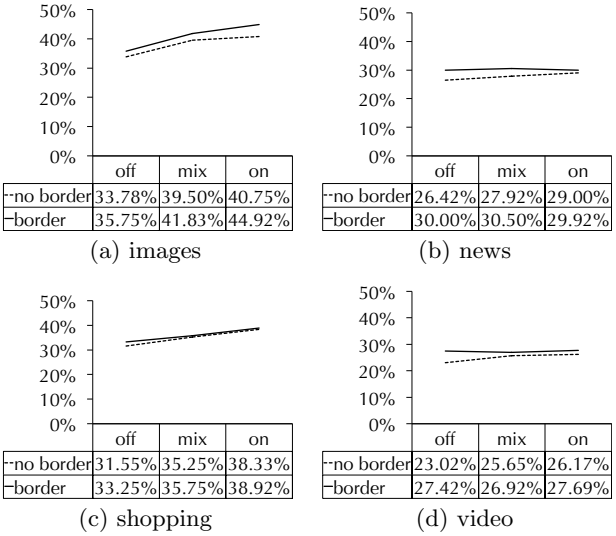


Figure 7: Interaction of border and sense on clicked web.

5. DISCUSSION

In the following sections, we discuss our results in terms of our three research questions (RQ1-RQ3).

5.1 Vertical Query-sense

Our first research question (RQ1) investigates whether the spill-over effect generalizes across verticals. The results from Section 4.1 suggest that the spill-over effect was stronger for images and shopping than for news and video. Images and shopping had significant effects for all four interaction measures, while news and video had significant effects for only scroll and mouseover, respectively. Arguello and Capra [2] investigated the same set of verticals and measured spill-over in terms of selected web and clicked web. They found significant spill-over effects for images and shopping, but not for news and video. Our results are consistent with theirs, but also show that news and video do have *some* level of spill-over, which can be detected using the lower-engagement signals derived from scrolls and mouseovers.

A natural question is: How are images and shopping different from news and video? To shed light on this, we consider how spill-over occurs and discuss differences in how the verticals were displayed on the SERP. Arguello and Capra [2] suggest that in order for spill-over to occur, the user must: (a) notice the vertical results, (b) process the vertical results and recognize their query-senses, and (c) allow the vertical results to influence their perception of the web results. As shown in Figure 2, our vertical surrogate representations varied across three dimensions: (1) the orientation of the vertical results, (2) the amount of text in the surrogate, and (3) the quality of the thumbnail image (i.e., the extent to which the thumbnail conveys the query-sense of the result). We conjecture that these three factors affected the vertical's visual salience on the SERP (affecting the likelihood of users noticing the vertical) and the level of cognitive effort required to process the vertical results (affecting the likelihood of users recognizing their query-senses).

Horizontal vs. vertical orientation—Image, shopping, and video results were displayed horizontally in a single row, making them visually distinct from the web results. In contrast, news results were displayed vertically, in a style similar to the web results. Models of visual processing indicate that people focus their visual attention on regions that exhibit 'pop-up' or distinguishability from the background [27]. In this respect, images, shopping, and video (which were more distinguishable from the web results) may have had more spill-over because they were noticed more by participants.

Amount of text—Image results were displayed using only image thumbnails. In contrast, news, shopping, and video results were displayed using a combination of thumbnails and other textual elements. Eye-tracking studies have found that when surrogates include both thumbnails and text, users focus on *both* elements. [14, 20].[3] These findings suggest that the image results (which included only thumbnails) required less cognitive effort to process, making it easier for participants to recognize their query-senses.

Image quality—Visual inspection suggests that the thumbnails used for images and shopping were more effective in conveying the underlying query-sense than those used for news and video. This point is illustrated in Figure 2. The thumbnails for the image vertical were obtained from the Bing Image Search API. In general, these are high-quality images that were selected by the search engine because they contain all the necessary information for judging relevance. The shopping thumbnails were obtained from the eBay Product Search API and mostly corresponded to photographs with a clear and sharp focus on the product. The news and video thumbnails were different. The news thumbnails were selected manually from the underlying news articles [2], but were considerably smaller ($\sim 50 \times 70$ pixels) than the thumbnails for the other verticals ($\sim 120 \times 120$ pixels). Finally, the video thumbnails were obtained from the Bing Video Search API and corresponded to video keyframes. The resolution, contrast, and focus of these keyframes seemed to vary more in quality. The pattern that emerges is that images and shopping used thumbnails that made it easier for participants to recognize the query-senses in the vertical results.

[3]Prior work suggests that users focus on the text first and then rely the thumbnail for confirmatory evidence [14].

Putting these factors together, we conjecture that the image vertical had the strongest spill-over because it was the most visually salient (distinguishable from the web results) and required the least cognitive effort (high-quality images with no text). On the other extreme, the news vertical had the weakest spill-over because it was the least salient (less distinguishable from the web results) and required the most cognitive effort (used significantly smaller thumbnails and the most text). Shopping and video both had display characteristics that placed them in-between images and news in terms of salience and cognitive effort. Thus, they had spill-over effects that were in-between images and news.

5.2 Vertical Rank and Query-Sense

Our second research question (RQ2) investigates whether vertical rank moderates the level of spill-over. Our primary interest in this question was to see if displaying the vertical at rank one causes more spill-over than displaying it at rank four. In addition, we were interested in understanding the main effect of rank, regardless of query-sense.

First, we consider whether the vertical rank moderates the level of spill-over (Figures 6(a)-6(d)). As discussed previously, in these figures, a spill-over is evidenced by an upward slope in the line as it moves from the off-target to the on-target condition (with the mixed condition in the middle). An upward slope indicates more interaction with the web results when the vertical results were more on-target. A steeper slope indicates a greater level of spill-over.

Three trends are worth noting. First, rank had a small moderating effect for news and video (not significant). As shown in Figures 6(b) and 6(d), news and video had virtually no spill-over at rank four (nearly flat solid lines) and a small level of spill-over at rank one (slightly upward slope for both dashed lines). Second, rank also had a small moderating effect for images (not significant). As shown in Figure 6(a), images had a high level of spill-over at rank four (steep upward slope for solid line) and a slightly *higher* level of spill-over at rank one (slightly steeper upward slope for dashed line). As shown in Figure 6(c), rank had a strong and statistically significant moderating effect for shopping ($p < .05$). Comparing off- versus on-target shopping results, the spill-over was about 10% at rank four (a change from 34.22% to 37.67%) and increased to about 30% at rank one (a change from 30.58% to 39.58%).

All verticals had more spill-over at rank one than rank two. However, these results suggest that rank has a stronger moderating effect for verticals with a mid-level of spill-over (e.g. shopping). In terms of RQ1, images had the strongest spill-over, and news and video had the weakest spill-over. In terms of RQ2, images had a strong spill-over irrespective of rank, and similarly, news and video had a weak spill-over irrespective of rank. Shopping appears to be at a middle point where rank can have a strong moderating effect. We see two possible explanations for this. One is that image results were highly salient and participants noticed and processed them regardless of rank. Another explanation is that after expending cognitive effort in processing the news and video results at rank one, participants were less willing to explore further down the results (even when the news/video results were on-target).

Since only the shopping vertical had a significant interaction effect, we were also interested in understanding the main effect of vertical rank on user interaction with the web results. The findings presented in Section 4.2 (and Figure 4) show that, in general, user interaction with the web results was lower when the vertical was presented at rank one versus rank four, regardless of query-sense. This trend suggests that participants scanned the SERP from top-to-bottom and did *not* skip-over the vertical when it was presented at rank one. Instead, we believe that participants expanded effort on the vertical results, which resulted in less interaction with the web results. While this trend was the same for all verticals, it was only statistically significant for news and video. One possible explanation is that when news and video were presented at rank one, users expended more cognitive effort on them (as compared to images and shopping) and were soon-after ready to explore a different query. This is consistent with our explanation for why the moderating effect of vertical rank was not significant for news and video.

5.3 Vertical Border and Query-Sense

Our third research question (RQ3) investigates whether including a border around the vertical results moderates the level of spill-over from the vertical to the web results. Our primary interest in this question was to see if the presence of a border helps signal to users that the vertical results are separate and distinct from the web results, resulting in less spill-over. In addition, we were interested in understanding the main effect of border, regardless of query-sense.

First, we consider whether including a border around the vertical results moderates the level of spill-over (Figures 7(a)-7(d)). None of the verticals had a significant interaction effect of vertical border and query-sense. However, we note several interesting trends in the data that give insights into the role that border played. First, for news and video (the ones with the least overall spill-over), border had a noticeable effect. Both verticals had virtually no spill-over in the presence of a border (indicated by the nearly flat solid lines in Figures 7(b) and 7(d)), but had a noticeable spill-over in the absence of a border (indicated by the upward trending dashed lines). Second, for shopping (mid-level of spill over), border had almost no effect (the solid and dashed lines are almost identical). Finally, for images (the one with the most overall spill-over), the border had a small, but opposite effect. In this case, the presence of a border increased the level of spill-over (the slope of the solid line is more than the slope of the dashed line).

These results suggest that including a border may reduce the level of spill-over more for those verticals that are less visually salient and require more cognitive effort to process (e.g. news and video). For these verticals, the border may play a role in allowing users to identify the vertical results as a group and shift their focus to the web results.

While we found some interesting trends, the moderating effect of border was not strong enough to reach significance for any vertical. It may be that the Gestalt principle of similarity (similar items are grouped together) also played a role. Our vertical results were visually distinct from the web results and this may have influenced participants to view them as a group regardless of a border. Future work might investigate the moderating effect of border for verticals that are closer in appearance to the web results.

Since we did not find strong significant interaction effects, we also explored the main effect of border on user interaction with web results. The findings presented in Section 4.3 (and Figure 5) show that, in general, user interaction with the web

results was greater when the vertical results were enclosed in a border. This effect was statistically significant ($p < .05$) for images, news, and video, and the trend was similar (but not significant) for shopping. One possible explanation is that the presence of a border helped participants to identify the vertical results as being a different type of result than the web results. The border might have allowed participants to skip-over the vertical results and focus more cognitive resources on processing the web results.

6. CONCLUSION

We reported on a large-scale user study that investigated whether the spill-over effect (from a set of vertical results to the web results) generalizes across verticals (RQ1), and whether the vertical rank (RQ2) and the presence of a border around the vertical results (RQ3) moderates this effect. Our results suggest three important trends. First, the spill-over effect generalizes across verticals, but is stronger for some verticals than others. Images had the strongest spill-over, followed by shopping, and then news and video. We confirm previous results that also found a stronger spill-over for images and shopping [2], but also show that news and video have *some* level of spill-over, which can be measured using interaction signals derived from scrolls and mouseovers. We conjecture that the images and shopping verticals in our study had more spill-over because they were more visually salient and required less cognitive effort to process.

Second, our results suggest that vertical rank has a stronger moderating effect for verticals with a mid-level of spill-over (e.g., shopping). In other words, verticals with a strong spill-over (e.g., images) or a weak spill-over (e.g., news and video) will have a similar level of spill-over irrespective of rank. A possible explanation is that highly salient verticals (e.g., images) are more likely to be noticed at lower ranks, and that verticals that require more cognitive effort (e.g., news and video) may influence users to re-formulate the query even when the vertical results are on-target.

Finally, including a border around the vertical results has a subtle effect on those verticals with a weak level of spill-over (e.g., news and video). A possible explanation is that a border influences users to skip-over the vertical results when they are less salient and require more cognitive effort.

Our findings have important implications for aggregated search. Current approaches to aggregated search do not *explicitly* favor coherent results. However, our research shows that the vertical results can influence user interaction with *other* components on the SERP. Moreover, the level of influence may depend on the specific vertical, its surrogate representation, where it is displayed, and how it is distinguished in the layout from other components on the SERP. Areas for future research include developing evaluation methodologies and algorithms that model all these different factors.

7. REFERENCES

[1] J. Arguello and R. Capra. The effect of aggregated search coherence on search behavior. In *CIKM*, pages 1293–1302. ACM, 2012.

[2] J. Arguello, R. Capra, and W.-C. Wu. Factors affecting aggregated search coherence and search behavior. In *CIKM*, pages 1989–1998. ACM, 2013.

[3] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. In *CIKM*, pages 201–210. ACM, 2011.

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

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

[6] P. Bailey, N. Craswell, R. W. White, L. Chen, A. Satyanarayana, and S. M. Tahaghoghi. Evaluating search systems using result page context. In *IIiX*, pages 105–114. ACM, 2010.

[7] G. Buscher, S. T. Dumais, and E. Cutrell. The good, the bad, and the random: An eye-tracking study of ad quality in web search. In *SIGIR*, pages 42–49. ACM, 2010.

[8] C. W. Cleverdon. The aslib cranfield research project on the comparative efficiency of indexing systems. *Aslib Proceedings*, 12(12):421–431, 1960.

[9] E. Cutrell and Z. Guan. What are you looking for?: An eye-tracking study of information usage in web search. In *CHI*, pages 407–416. ACM, 2007.

[10] F. Diaz. Integration of news content into web results. In *WSDM*, pages 182–191. ACM, 2009.

[11] F. Diaz and J. Arguello. Adaptation of offline vertical selection predictions in the presence of user feedback. In *SIGIR*, pages 323–330. ACM, 2009.

[12] S. T. Dumais, G. Buscher, and E. Cutrell. Individual differences in gaze patterns for web search. In *IIiX*, pages 185–194. ACM, 2010.

[13] G. Hotchkiss. Q&A with Marissa Mayer. http://searchengineland.com/qa-with-marissa-mayer-google-vp-search-products-user-experience-10370. Accessed: 2014-06-01.

[14] A. Hughes, T. Wilkens, B. M. Wildemuth, and G. Marchionini. Text or pictures? an eyetracking study of how people view digital video surrogates. In *CIVR*, pages 271–280. Springer-Verlag, 2003.

[15] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161. ACM, 2005.

[16] S. Kalyanaraman and J. D. Ivory. Enhanced information scent, selective discounting, or consummate breakdown: The psychological effects of web-based search results. *Media Psychology*, 12(3):295–319, 2009.

[17] G. Keppel and T. D. Wickens. *Design and Analysis: A Researcher's Handbook*. Prentice Hall, 3 edition, 1991.

[18] K. Koffka. *Principles of Gestalt psychology*. Harcourt, New York, 1935.

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

[20] H. A. A. Maqbali, F. Scholer, J. Thom, and M. Wu. Evaluating the effectiveness of visual summaries for web search. In *ADCS*, pages 36–43, 2010.

[21] S. E. Palmer. Common region: A new principle of perceptual grouping. *Cognitive Psychology*, 24(3):436 – 447, 1992.

[22] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

[23] A. K. Ponnuswami, K. Pattabiraman, D. Brand, and T. Kanungo. Model characterization curves for federated search using click-logs: predicting user engagement metrics for the span of feasible operating points. In *WWW*, pages 67–76. ACM, 2011.

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

[25] M. Sanderson. Ambiguous queries: test collections need more sense. In *SIGIR*, pages 499–506. ACM, 2008.

[26] R. L. T. Santos, C. Macdonald, and I. Ounis. Aggregated search result diversification. In *ITCIR*, pages 250–261. Springer-Verlag, 2011.

[27] F. Stentiford. Visual attention: low-level and high-level viewpoints. *SPIE*, 8436:84360L–84360L–9, 2012.

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

[29] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, CA, USA, 2000.

[30] K. Zhou, R. Cummins, M. Lalmas, and J. M. Jose. Evaluating aggregated search pages. In *SIGIR*, pages 115–124. ACM, 2012.