The Effect of Aggregated Search Coherence on Search Behavior

Jaime Arguello
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599-3360 USA
jarguello@unc.edu

Robert Capra
School of Information and Library Science
University of North Carolina at Chapel Hill
Chapel Hill, NC, 27599-3360 USA
rcapra@unc.edu

ABSTRACT

Aggregated search is the task of blending results from different specialized search services, or *verticals*, into the web search results. Aggregated search coherence refers to the degree to which results from different systems focus on similar senses of the query. While cross-component coherence has been cited as an important criterion for whole-page evaluation, its effect on search behavior has not been deeply investigated in prior research. In this work, we focus on the coherence between two aggregated search components: images and web results. In particular, we investigate whether the query-senses associated with the blended image results can influence user interaction with the web results. For example, if a user wants web results about "jaguar" the animal, are they more likely to examine the web results if the image results contain pictures of the animal instead of pictures of the car? Based on two large user studies, our results show that the image results can systematically affect user interaction with the web results. If the web results are largely consistent with the search task, then the effect of the image results is small. However, if the web results are only marginally consistent with the search task, such as when they are highly diversified across query-senses, the image results have a significant effect on user interaction with the web results. Our findings have implications on current research in whole-page evaluation, aggregated search, and diversity ranking.

Categories and Subject Descriptors

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

General Terms

Performance, Experimentation, Human Factors

Keywords

Aggregated search, aggregated search coherence, user study, evaluation, search behavior, assimilation effects

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$10.00.

1. INTRODUCTION

Commercial search services (e.g., Bing, Google, and Yahoo!) currently provide access to a wide range of systems, known as *verticals*. Example verticals include search engines for images, videos, and news articles, as well as applications that provide information such as weather forecasts, stock quotes, and driving directions. While many verticals have direct search capabilities, in some situations a user may not know that a vertical is relevant or may want results from several verticals at once. Thus, an increasingly important problem for commercial search services is the prediction and integration of relevant vertical results into the core web results. In the research literature, this task is referred to as *aggregated search.* Typically, verticals are presented by blending a few of their results somewhere above, within, or below the first page of web results. The end goal is to either satisfy the user directly with the aggregated vertical results, or to at least convey how the information need might be satisfied by searching within a specific vertical.

Aggregated search coherence refers to the extent to which the results from different systems (different vertical search engines and the web search engine) focus on similar senses of the query. Suppose that a user issues the query "joplin" and the aggregated system decides to blend image results into the web results. If the web results focus on the town of Joplin, Missouri and the blended image results focus on the artist Janis Joplin, then one might say that the web and image results have a *low* level of coherence. Conversely, if both types of results focus on the same query sense (the town or the artist) or a similar mixture of both, then one might say that the web and image results have a *high* level of coherence. In this work, we investigate the effect of aggregated search coherence on search behavior. If a user is looking for web results about Joplin, Missouri and the blended image results are mostly pictures of Janis Joplin, does this affect the user's decision to explore the web results?

Traditional IR evaluation uses Cranfield-style methodologies involving a corpus, a set of queries, relevance judgments made outside the context of a particular retrieval, and evaluation metrics that operate on a ranking of judged documents [11]. In recent years, however, new evaluation methods have been proposed in order to address the fact that users interact with results within the context of other results and within the context of other information displayed in the results page (web results, vertical results, query suggestions, advertisements, etc.)[6]. While cross-component coherence has been cited as an important criterion for whole-page evaluation, its effect on users' assessments of the search results, or its effect on search behavior, has not been investigated in prior work and many questions remain. Can the blended results from a particular vertical affect a user's evaluation of the web results and their decision to interact with them? Are visually salient verticals like *images* more influential than textual verticals like *news*? Does it depend on where the vertical is blended? Does it also depend on the web results?

As a starting point into these research questions, the current paper focuses on two aggregated search components: images and web results. We present two user studies that investigate how the blended image results can affect a user's assessment of the web results and, consequently, their decision to interact with the web results or reformulate the query. More concretely, we explore the following scenario. A user wants web results on a particular topic (e.g., tourism information for Joplin, Missouri) and enters an underspecified, ambiguous query (e.g., "joplin"). In response to this query, the system decides to blend image results into the web results. Is the user more likely to interact with the web results if the image results are consistent with the search task (i.e., pictures of Joplin, Missouri vs. pictures of Janis Joplin)? And, if the image results have an effect, does the magnitude of the effect also depend on the web results?

Several trends suggest that the scenario we explore is a practical problem. First, queries are oftentimes ambiguous. Sanderson [22] found that about 4% of all unique queries and 16% of all head queries issued to a commercial search engine are ambiguous. Second, given an ambiguous query, the distribution of senses represented in the top results from different systems may be different for two reasons: the distribution of senses within different document collections may be different (e.g., an image collection might have more images of Janis Joplin than Joplin, Missouri) and the distribution of senses associated with queries issued to different systems may be different (e.g., a news vertical may see more users interested in Joplin, Missouri than Janis Joplin).[1] Given that search engine results are influenced by the content in the collection as well as previous user interactions, it is conceivable for two different systems to retrieve a different distribution of senses for the same query. Finally, previously published aggregated search techniques are not explicitly designed or tuned to ensure coherence between results from different systems. While most aggregated search techniques use machine learning to combine different types of features [4, 5, 2, 13, 20, 19], none of the previously reported features focus on coherence. Furthermore, coherence is not explicitly considered in existing methods for aggregated search optimization [4, 5, 2, 3, 13, 20, 19].

Outside of the field of information retrieval, an extensive amount of research has shown that in certain situations people associate attributes of a contextual stimulus to an object being judged. This effect is known as an *assimilation* effect and has been observed, for example, in people's judgements about the quality of a business [17], the quality of a product [24, 18], or the meaning of a survey question [12, 26]. To our knowledge, assimilation effects have not been investigated in prior work in information retrieval.

Within the current study, we treat the blended image results as the contextual stimulus and the web results as the object being judged. Our main hypothesis is that if the blended images are on the same sense as the search task, an assimilation effect causes users to make a more positive preliminary assessment of the web results, as indicated by their decision to interact with the web results. Conversely, if the blended images are on a different sense as the search task, an assimilation effect causes users to make a more *negative* preliminary assessment of the web results, as indicated by their decision to reformulate the query without interacting with the web results.

1.1 Research Questions

We study the effects of aggregated search coherence on search behavior and focus on two specific aggregated search components: image vertical results and web results. In particular, we investigate whether the query-senses represented within the blended image results can affect user interaction with the web results, and, if so, whether the magnitude of the effect depends on the web results. More explicitly, we address the following research questions.

RQ1: How do the senses represented in the web results affect user interaction with the web results?

RQ2: How do the senses represented in the blended image results affect user interaction with the web results?

RQ3: Are there interaction effects? Does the effect of the images also depend on the web results?

RQ4: How do the senses represented in the image results affect user interaction with the types of web results returned by a competitive commercial system?

2. RELATED WORK

Query ambiguity has long been recognized as an important problem in information retrieval. If a user issues the query "jaguar", do they mean the animal or the car? A recent query-log analysis conducted by Sanderson [22] found that about 4% of all unique queries and about 16% of all unique head queries issued to a commercial search engine had multiple senses. In the current study, we investigate whether the query-senses represented in the blended image results affect user interaction with the web results. Sanderson's analysis shows that ambiguous queries are common.

Given an ambiguous query, a common strategy for search systems is to *diversify* the search results (e.g., to return results about jaguar the car *and* the animal). The importance of diversifying search results is reflected in the development of new algorithms for ranking [28, 21, 7] and metrics for evaluation [9, 1, 8, 10]. Common to all these methods, however, is the assumption of *homogeneous* results (e.g., only web results). None of these methods for diversity ranking and evaluation consider the potential impact of vertical results being blended into the web results (particularly from visually salient verticals like *images*).

Aggregated search is the task of blending vertical results into the web results and is typically decomposed into two tasks: predicting *which* verticals to present (*vertical selection* [4, 5, 13, 15, 16]) and predicting *where* in the web results to present them (*vertical presentation* [2, 3, 20, 19]). Existing methods for vertical selection and presentation use machine learning to combine different types of features: properties of the query string [4, 5, 2, 15], the predicted relevance of the vertical results [4, 5, 2, 13, 15], the similarity between

[1] The U.S. town of Joplin, Missouri has been in the news since it was hit by a tornado in 2011.

the query and those issued directly to the vertical by users [4, 5, 13], and implicit feedback signals derived from previous presentations of the vertical [13, 20, 19]. None of the previously reported features, however, explicitly consider the coherence between results from different sources.

Aggregated search evaluation is typically done with respect to editorial judgments—a human assessor decides that a particular vertical is relevant to the query and should be blended in a particular position [4, 5, 14, 3]—or, in a production environment, with respect to user-generated clicks and skips [13, 20]. Both evaluation methods punish false-positive vertical predictions. One of our contributions is that we investigate whether certain false-positive predictions degrade the user experience more than others. For example, suppose a user only wants web results and the system decides to present image results (i.e., arguably a false-positive error). Can certain image results have a greater effect on the user's assessment of the web results?

A few user studies have investigated search and preference behavior with aggregated search interfaces. These studies have shown two important trends. First, users prefer to see results from relevant verticals at the top of the blended results [25, 27] and, second, users tend to click more on visually salient verticals (e.g., *video*) irrespective of rank and relevance [25]. We build upon this work by investigating the effect of results from the images vertical (also visually salient) on user interaction with results from a *different* source (i.e., the web results).

Search engines present users with many different types of information (e.g., web results, vertical results, query suggestions, related searches, advertisements, etc.). Bailey *et al.* propose a whole-page evaluation methodology referred to as *Student Assignment Satisfaction Index* (SASI) [6]. The evaluation methodology focuses on eliciting quality judgments from assessors on parts of the SERP within the context of the whole. While query-sense coherence between components is mentioned as an important criterion for whole-page evaluation, its effect on user behavior or user satisfaction was not investigated [6].

Outside of information retrieval, the effect of images on people's responses to textual elements has been investigated within the field of survey design. While survey designers sometimes use pictures to supplement survey questions and make surveys more engaging, prior research shows that pictures can systematically affect respondents' interpretation of a survey question [12, 26]. Couper *et al.* investigated the effect of pictures on survey questions that asked respondents how frequently they engaged in a specific activity (e.g., going shopping or eating out) [12]. Respondents exposed to a "high-frequency" image (e.g., an image of someone shopping for groceries or eating fast-food) reported greater numbers than those exposed to a "low-frequency" image (e.g., an image of someone shopping for clothes or eating at an intimate restaurant). The pictures influenced the respondents' interpretation of the text. In more recent work, Toepoel and Couper found that the influence of pictures can be counteracted by making the survey question more explicit (e.g., by describing what qualifies as "going shopping" or "eating out", whether it be a high- or low-frequency definition) [26].

Finally, and more generally, within the fields of psychology and marketing, a significant amount of research has been aimed at understanding how contextual stimuli (e.g., an image) can affect a person's assessment of an object (e.g., a survey question) [23]. One such effect is referred to as an *assimilation* effect—a person assigns attributes of the contextual stimulus to the object being evaluated. In psychology studies, an assimilation effect is usually measured based on subjects' responses to specific questions about the object being evaluated. For example, Meyers *et al.* [17] asked participants to rate a new restaurant after being told that the restaurant's venue was previously occupied by a different high-end restaurant (in one condition) vs. a fast-food restaurant (in another condition). As predicted, subjects in the first condition rated the restaurant more favorably. In the current study, we treat the blended image results as the contextual stimulus and the web results as the object being evaluated. However, we did not ask participants to *explicitly* rate the web results. Instead, we observed their search behavior under different conditions. Our assumption is that participants did not interact with the web results when they believed the web results did not contain the sought-after information or when they did not want to expend the effort to make a more thorough assessment.

3. METHOD

Two separate studies were conducted, referred to as Study 1 and Study 2. Both studies consisted of subjects completing a series of search tasks using a live search engine and both studies had a similar objective: to investigate whether the query-senses represented in the blended image results have an effect on users' interaction with the web results.

3.1 Experimental Protocol

Study 1 and Study 2 followed a similar protocol (Figure 1). Subjects were given access to a live search engine and asked to find a webpage containing a specific type of information. Search task descriptions were given in the form of a question (e.g., "What country makes the electric sports car Tesla?") or a request for information (e.g., "Find a website that contains tourism information about Santa Fe, New Mexico."). Prior to starting the search session, subjects were told that in order "to get them started" they would be presented with an initial query and a set of results. We refer to this as the initial SERP. From the initial SERP, subjects were instructed to search naturally, by either examining the results provided in the initial SERP or by issuing their own queries. Selecting a particular webpage as "containing the requested information" concluded the task. The live search engine provided to participants was implemented using the Bing Web Search API. All queries, results, clicks, and webpage selections were recorded in our server.

In both studies, the initial SERP is where the experimental manipulation took place. The initial SERP was artificially constructed in order to test different experimental conditions, for example, by blending images into the web results and varying the distribution of senses represented in the image results. The experimental manipulation of the initial SERP was not communicated to the study participants. Outcome measures were derived from subjects' interactions with the initial SERP. For example, we considered the number of times a participant clicked on a web results in the initial SERP. The initial SERP was the only SERP that was artificially altered. All user-generated queries were issued to the Bing Web Search API in real-time and the actual algorithmic results were presented.

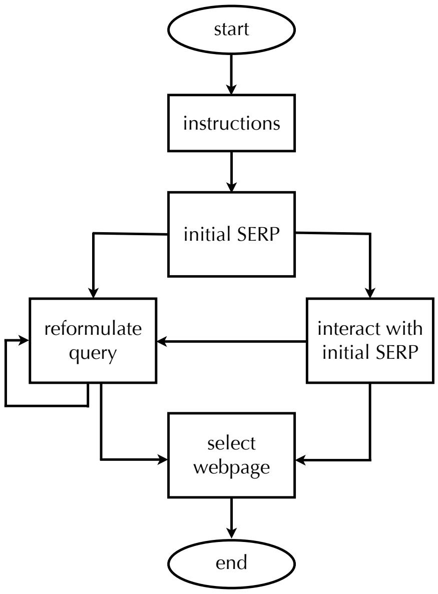


Figure 1: Flow diagram description of the protocol used in both studies.

Both studies were run using Amazon's Mechanical Turk (AMT).[2] Amazon's Mechanical Turk is a crowdsourcing internet marketplace in which *requesters* can publish relatively simple tasks, referred to as Human Intelligence Tasks (HITs), to be completed by *workers* in exchange for compensation. Search tasks were implemented as *external* HITs, meaning that all interactions with our search system, including the assignment of subjects (or workers in AMT parlance) to search tasks and experimental conditions, was managed in our own server. As described in more detail later, this allowed us to avoid learning effects and to do early detection and filtering of careless workers.

Figure 2 shows a few screenshots from one of our HITs. At the start of each HIT, subjects were first given a set of instructions (Figure 2(a)). A hyperlink provided in these instructions took participants to a more detailed description of the protocol, which included a description of the initial SERP as a means to "get them started" with the search task. After reading these instructions, participants were asked to click the "start task" button in order to begin the search task. Clicking the "start button" opened the initial SERP in a separate browser tab (Figure 2(b)). The initial SERP displayed the search task description at the top (in this case, "Find the official website of the Irish musical band the Cranberries."), a task-specific query (in this case, "cranberries"), and a set of web results along with a set of optional blended image results (in this experimental condition, pictures of cranberries the fruit). The search task description was provided as an embedded image to prevent participants from copying and pasting it into the query box. From the initial SERP, subjects were free to search naturally by either examining the results provided in the initial SERP or by issuing their own queries (Figure 2(c)). All user-generated queries were issued to the Bing Web Search API (via our server). Clicking on a search result (either a web result presented in the initial SERP or one returned in response to a user-generated query) displayed the webpage in an HTML frame covering almost the entire screen (Figure 2(d)). Above this frame, we provided the search task description and a button labeled "Click here if the this pages contains the requested information". Clicking this button selected the current page as containing the answer to the question or the requested information. Finally, participants were given a task-specific completion code (Figure 2(e)) and were asked to copy and paste this code into the "validate code" textbox in the original instructions page (Figure 2(f)). After validating the completion code entered, participants were allowed to submit the HIT.

3.2 Search Tasks

Each of our search tasks consists of four components. The *search task description* corresponds to a question or request for information. Subjects were given the search task description and asked to find a webpage containing the requested information (e.g., "Find a website that lists places in the world where pumas can be found in the wild."). Additionally, each search task had two senses: the *target sense* corresponds to the same sense as the search task description (e.g., puma the animal) and the *off-target sense* corresponds to a different sense (e.g., Puma the athletic shoe brand). Finally, the *initial query* corresponds to a short string that a user might issue to a search engine while looking for the information requested in the search task description (e.g., "pumas"). The initial query, as opposed to the search task description, was ambiguous in terms of the desired sense: the target sense (e.g., puma the animal), the off-target sense (e.g., Puma the athletic shoe brand), or any other sense (e.g., Puma the Brazilian sports car).

A total of 105 search tasks were created using the following procedure. First, it was necessary to identify a set of ambiguous entities (single terms or phrases with multiple meanings). Following Sanderson [22], we collected a large set of ambiguous entities by automatically identifying all the Wikipedia disambiguation pages. In Wikipedia, a disambiguation page serves as a navigational hub, providing links to Wikipedia articles on specific senses of the entity.[3] Wikipedia disambiguation pages were identified using regular expressions (e.g., pages having "(disambiguation)" in the title or the "{{disambig}}" tag in the Wiki markup). A total of 122,130 Wikipedia disambiguation pages were identified.[4]

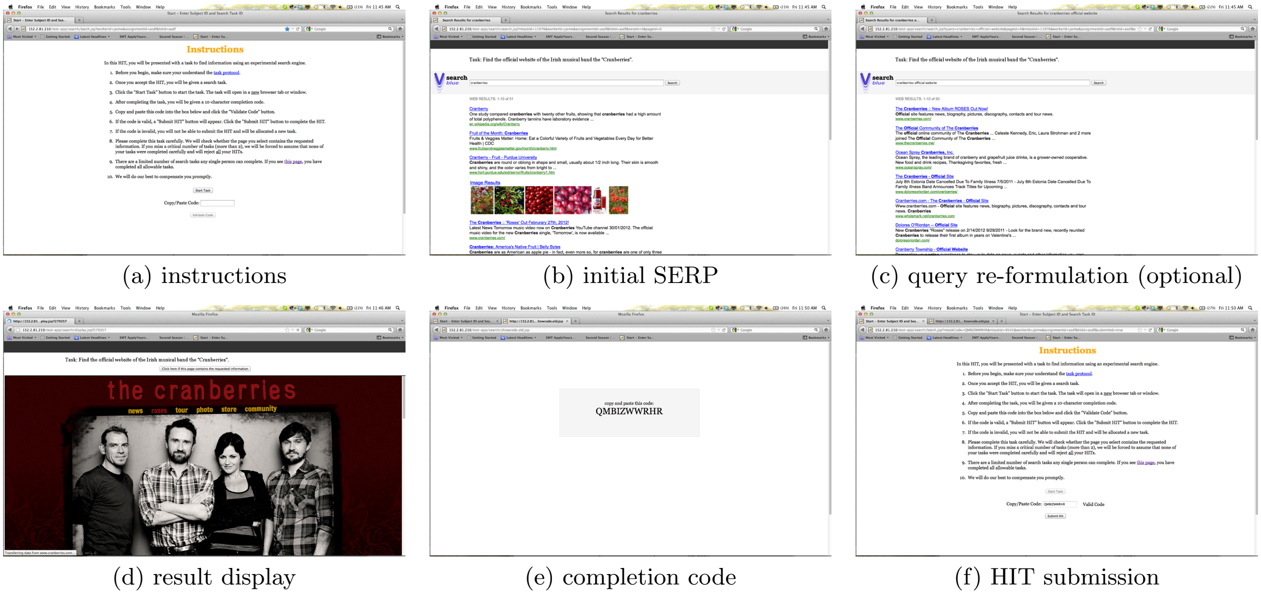
The second step was to identify the subset of ambiguous entities that might be issued as a query to a commercial system. We did not want participants to be surprised by seeing the entity as the initial query. To this end, we omitted all ambiguous entities not appearing in the AOL query-log. Out of the original 122,130 ambiguous entities, 34,151 (28%) had an exact-match query in the AOL query-log.

The third step was to identify those entities with exactly two senses with a strong image orientation. We did not want participants to be surprised by seeing blended image results in response to the entity, and more specifically, image results associated with a mixture of senses (e.g., pictures of pumas the animal *and* Puma shoes). In other words, we wanted to avoid entities such as "leopard", which has two senses (the animal and the operating system), but only one sense (the animal) has a strong image orientation. This goal was

[2] `https://www.mturk.com`

[3] See, for example, `http://en.wikipedia.org/wiki/Puma`

[4] We believe this number is larger than that reported in Sanderson [22] because we used a newer version of Wikipedia (latest available on 11/19/2011).


Firefox File Edit View History Bookmarks Tools Window Help
Most Visited Getting Started Latest Headlines EMT ApplyYours... Second Season | ... Start – Enter Su... Bookmarks
Google

Start – Enter Subject ID and Search Task ID

152.2.81.210/test-app/search/batch.jsp?workerId=jaime&assignmentId=asdf&hitId=asdf

Instructions

In this HIT, you will be presented with a task to find information using an experimental search engine.

1. Before you begin, make sure your understand the task protocol.
2. Once you accept the HIT, you will be given a search task.
3. Click the "Start Task" button to start the task. The task will open in a new browser tab or window.
4. After completing the task, you will be given a 10-character completion code.
5. Copy and paste this code into the box below and click the "Validate Code" button.
6. If the code is valid, a "Submit HIT" button will appear. Click the "Submit HIT" button to complete the HIT.
7. If the code is invalid, you will not be able to submit the HIT and will be allocated a new task.
8. Please complete this task carefully. We will check whether the page you select contains the requested information. If you miss a critical number of tasks (more than 2), we will be forced to assume that none of your tasks were completed carefully and will reject all your HITs.
9. There are a limited number of search tasks any single person can complete. If you see this page, you have completed all allowable tasks.
10. We will do our best to compensate you promptly.

Start Task

Copy/Paste Code:

Validate Code

(a) instructions

Search Results for cranberries

152.2.81.210/test-app/search/search.jsp?mtaskId=11976&workerId=jaime&assignmentId=asdf&hitId=asdf&serpId=0&pageId=0

Task: Find the official website of the Irish musical band the "Cranberries".

search blue | cranberries | Search

WEB RESULTS: 1-10 of 51

Cranberry
One study compared **cranberries** with twenty other fruits, showing that **cranberries** had a high amount of total polyphenols. Cranberry tannins have laboratory evidence ...
en.wikipedia.org/wiki/Cranberry

Fruit of the Month: **Cranberries**
Fruits & Veggies Matter: Home: Eat a Colorful Variety of Fruits and Vegetables Every Day for Better Health | CDC
www.fruitsandveggiesmatter.gov/month/cranberry.html

Cranberry - Fruit - Purdue University
Cranberries are round or oblong in shape and small, usually about 1/2 inch long. Their skin is smooth and shiny, and the color varies from bright to ...
www.hort.purdue.edu/ext/senior/fruits/cranberry1.htm

Image Results

The **Cranberries** :: 'Roses' Out Februrary 27th, 2012!
Latest News Tomorrow music video now on **Cranberries** YouTube channel 30/01/2012. The official music video for the new **Cranberries** single, 'Tomorrow', is now available ...
www.cranberries.com/

Cranberries: America's Native Fruit | Belly Bytes
Cranberries are as American as apple pie - in fact, even more so, for **cranberries** are one of only three

(b) initial SERP

Search Results for cranberries official website

152.2.81.210/test-app/search/search.jsp?query=cranberries+official+website&pageId=0&mtaskId=11976&workerId=jaime&assignmentId=asdf&hitId=asdf&s

Task: Find the official website of the Irish musical band the "Cranberries".

search blue | cranberries official website | Search

WEB RESULTS: 1-10 of 50

The **Cranberries** :: New Album ROSES Out Now!
Official site features news, biography, pictures, discography, contacts and tour news.
www.cranberries.com/

The **Official** Community of The **Cranberries**
The **official** online community of The **Cranberries** ... Celeste Kennedy, Eric, Laura Strohman and 2 more joined The **Official** Community of The **Cranberries** ...
www.thecranberries.me/

Ocean Spray **Cranberries**, Inc.
Ocean Spray, the leading brand of cranberry and grapefruit juice drinks, is a grower-owned cooperative. New food and drink recipes, Thanksgiving favorites, fresh ...
www.oceanspray.com/

The **Cranberries** - **Official** Site
July 8th Estonia Date Cancelled Due To Family Illness 7/5/2011 - July 8th Estonia Date Cancelled Due To Family Illness Band Announces Track Titles for Upcoming ...
www.doloresoriordan.com/cranberries/

Cranberries.com - The **Cranberries** - **Official** Site
Www.cranberries.com - **Official** site features news, biography, pictures, discography, contacts and tour news. **Cranberries**
www.whoismark.net/cranberries.com

Dolores O'Riordan -- **Official** Site
New **Cranberries** "Roses" release on 2/14/2012 9/29/2011 - Look for the brand new, recently reunited **Cranberries** to release their first album in years on Valentine's ...
doloresoriordan.com/

Cranberry Township - **Official Website**

(c) query re-formulation (optional)

Mozilla Firefox

152.2.81.210/test-app/search/display.jsp?279357

Task: Find the official website of the Irish musical band the "Cranberries".

Click here if this page contains the requested information

the cranberries
news roses tour photo store community
Transferring data from www.cranberries.com...

(d) result display

Mozilla Firefox

152.2.81.210/test-app/search/showcode.old.jsp

copy and paste this code:
QMBIZWWRHR

(e) completion code

Start – Enter Subject ID and Search Task ID

152.2.81.210/test-app/search/batch.jsp?mtaskCode=QMBIZWWRHR&mtaskId=9542&workerId=jaime&assignmentId=asdf&hitId=asdf&submitted=true

Instructions

In this HIT, you will be presented with a task to find information using an experimental search engine.

1. Before you begin, make sure your understand the task protocol.
2. Once you accept the HIT, you will be given a search task.
3. Click the "Start Task" button to start the task. The task will open in a new browser tab or window.
4. After completing the task, you will be given a 10-character completion code.
5. Copy and paste this code into the box below and click the "Validate Code" button.
6. If the code is valid, a "Submit HIT" button will appear. Click the "Submit HIT" button to complete the HIT.
7. If the code is invalid, you will not be able to submit the HIT and will be allocated a new task.
8. Please complete this task carefully. We will check whether the page you select contains the requested information. If you miss a critical number of tasks (more than 2), we will be forced to assume that none of your tasks were completed carefully and will reject all your HITs.
9. There are a limited number of search tasks any single person can complete. If you see this page, you have completed all allowable tasks.
10. We will do our best to compensate you promptly.

Start Task

Copy/Paste Code: QMBIZWWRHR Valid Code

Submit Hit

(f) HIT submission

Figure 2: Screenshots from one of our Mechanical Turk HITs.

accomplished in two steps. First, we issued each entity in the set of 34,151 to the Bing search engine and, via screen-scraping, identified the subset that triggered blended image vertical results. A subset of 3,452 entities (10%) triggered blended image results. Then, we used the Bing Images API to cache the top-20 image results returned in response to these entities. Topics were then randomly sorted and the first 105 entities with image results associated with exactly two senses were identified manually.

Finally, search tasks were created by constructing the search task description consistent with one of the two senses represented in the top-20 image results (e.g., "Find a website that lists places in the world where pumas can be found in the wild."). The target sense was set to the same sense as the search task description (e.g., puma the animal) and the off-target sense was set to the other sense represented in the top-20 image results (e.g., Puma the athletic shoe brand). The initial query was set to the Wikipedia entity (e.g., "puma"). Table 1 shows ten example search tasks.

3.3 Study 1

Study 1 used a full factorial design with three independent variables: *search task*, *blended images*, and *answer rank*. We used a total of 80 experimental search tasks designed as described above. The *blended images* variable manipulated the distribution of senses associated with the image results presented in the initial SERP and had four possible values:

- No images: no images blended into the initial SERP;
- On-target images: 6 images blended into the initial SERP between web results 3 and 4, with all images associated with the target sense (consistent with the search task description);
- Off-target images: 6 images blended into the initial SERP between web results 3 and 4, with all images associated with the off-target sense (inconsistent with the search task description); and
- Mixed images: 6 image results blended into the initial SERP between web results 3 and 4, with 3 images associated with the target sense and 3 images associated with the off-target sense.

The 12 images associated with each task (6 on-target and 6 off-target) were identified manually from the top-20 images returned by the Bing Image Search API and were cached in advance. Manual pruning was done in order to remove near-duplicates and to ensure good image quality. The 6 images displayed in the mixed images condition were selected randomly from the set of 12 (3 on-target and 3 off-target).

In addition to manipulating the blended image results, we also manipulated the web results. The *answer rank* variable was manipulated in the following manner. First, for each search task, we identified one web result containing the requested information. We refer to this web result as the *answer* page. Second, we identified ten web results returned in response to the initial query that were all associated with the off-target sense (inconsistent with the search task description). All web results (titles, summary snippets, and URLs) were cached in advance using the Bing Web Search API. The *answer rank* variable had five values: the answer page presented at rank 1, 3, 4, 8, and 10. The remaining 9/10 web results (excluding the answer page) were all on the off-target sense. The initial SERP layouts associated with Study 1 are shown Figure 3

Study 1 had 1,600 unique conditions (80 search tasks × 4 image conditions × 5 answer rank conditions = 1,600). Additionally, we collected six redundant data points per condition, for a total 9,600 experimental units. Each experimental unit was executed as a Mechanical Turk HIT. Each HIT was priced at $0.10 USD.

Running user studies on Mechanical Turk requires quality control. Careless workers were filtered in two ways. First, we restricted our HITs to workers with a 95% acceptance rate or greater and workers within the U.S. (to help ensure English proficiency). Second, we identified and filtered careless workers early. To this end, we used the fact that only

Table 1: Example search tasks. The search task description corresponds to the question or information request given to subjects, the target sense corresponds to the sense associated with the search task description, the off-target sense corresponds to a tangential, but still popular sense, and the initial query corresponds to the query displayed in the initial SERP.

Search Task Desciption	Target Sense	Off-Target Sense	Initial Query
Where in Europe are the Pyrenees Mountains located?	mountain range	dog breed	pyrenees
What is latest album released by Seal?	musician	animal	seal
Where in the U.S. is the island of Captiva?	island	automobile	captiva
Find a website that contains information about the Kiwi bird.	bird	fruit	kiwi
Find the official website of the SCUBA diving equipment manufacturer Aqualung.	company	musical band	aqualung
What is the real name of the U2 guitarist The Edge?	person	movie	the edge
Find a website that contains tourism information about Manhattan in NYC.	location	cocktail	manhattan
Find the official website of the musical band Rogue Wave	musical band	phenomenon	rogue wave
Is the lotus plant an aquatic plant?	plant	automobile	lotus
Do black sheep actually occur in nature?	animal	movie	black sheep

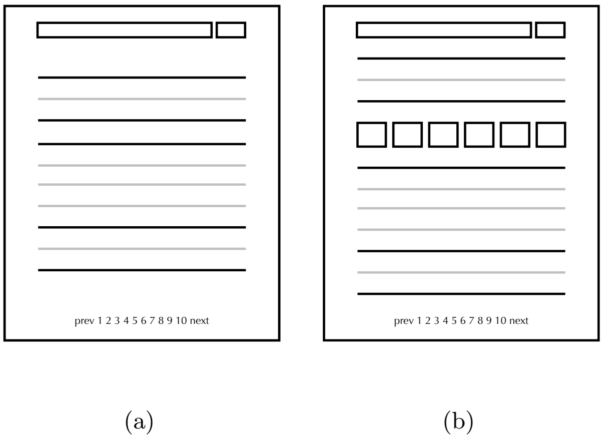

prev 1 2 3 4 5 6 7 8 9 10 next
prev 1 2 3 4 5 6 7 8 9 10 next
(a)
(b)

Figure 3: Study 1 layouts associated with (a) *blended images* = no images and (b) *blended images* = on-target, off-target, and mixed. The web ranks shown in black denote the possible ranks for the answer page.

one web results in the initial SERP was on the target sense. All other web results were on the off-target sense. We interpreted the selection of an off-target web result as "containing the requested information" as a sign of careless work. Participants who selected more than two off-target web results from the initial SERP were not allowed to do more HITs.

Finally, in order to avoid learning effects, the assignment of participants to experimental conditions was managed dynamically within our server, using the AMT Worker ID to identify the participant. Workers were assigned to experimental conditions based on the following criteria. First, workers were not exposed to the same search task more than once, even if the worker did not complete the task. Second, the combination of image condition and answer rank condition was randomized for each participant. Beyond these constraints, each participant was allowed to do as few or as many tasks as desired. Finally, to disguise the purpose of the study, we published 3,000 "distractor" HITs. These were associated with an additional 25 search tasks (different from the 80 experimental search tasks, for a total of 105 search tasks). The initial SERPs associated with distractor tasks did not present image results and presented the algorithmic web results exactly as returned by the Bing Web Search API (i.e., no experimental manipulation).

The web results in Study 1 were artificially altered, which has advantages and disadvantages. On one hand, it allowed us to perform a detailed analysis of search behavior. For example, it allowed us to determine whether users look more closely at the web result immediately above or immediately below the blended image results. On the other hand, the web results did not exactly match those returned by a competitive system for the initial query. We addressed this limitation in Study 2.

3.4 Study 2

Study 2 also had a full factorial design, but had only two independent variables: *search task* and *blended images*. The possible values associated with these variables were identical to those in Study 1. The difference was that in Study 2, the web results presented in the initial SERP were not manipulated. Instead, we always presented the top-10 algorithmic web results returned by the Bing Web Search API in response to the initial query. By doing so, we were able to study the effect of the blended image results on user interaction with state-of-the-art web results (**RQ4**).

While we did not manipulate the web results for each search task, we wanted to study the interaction effects between the query-sense distributions in the image and web results. To this end, we grouped search tasks into three bins: a *high*, *mid*, and *low* target-sense bin.

The binning was done as follows. First, for each search task, we manually identified which of the top-10 algorithmic web results were on the target sense (the same as the search task description). Second, for each search task, we scored the top-10 web results using NDCG@10, by considering the target-sense as the 'relevant' class. NDCG@10 was computed as,

$$\text{NDCG@10} = \frac{1}{\mathcal{Z}} \sum_{r=1}^{10} \frac{\text{sense}(r)}{\log_2(\max(r, 2))},$$

where $\mathcal{Z}$ is the the NDCG normalizing factor (assuming all top-10 web results on the target sense) and the function sense(r) returns 1 if the web result at rank r is on the target sense and 0 if it is on a different sense. A non-zero NDCG@10 score means that at least one web result is on the target sense. The greater the score, the greater the number of top-10 web results on the target sense and the higher their ranks. Finally, search tasks were binned by assigning the top 20% to the *high* bin (mean NDCG = 0.896, stdev = 0.095), the middle 60% to the *mid* bin (mean NDCG =

0.480, stdev = 0.138), and the bottom 20% to the *low* bin (mean NDCG = 0.156, stdev = 0.077).

Study 2 had 320 unique conditions (80 search tasks × 4 image conditions = 320). Additionally, we collected 20 redundant data points per condition, for a total 6,400 experimental units. Each experimental unit was executed as a Mechanical Turk HIT. Each HIT was priced at $0.10 USD.

Quality control was done the same as in Study 1. We interpreted the selection of an off-target web result in the initial SERP as evidence of careless work. Participants who selected more than two off-target web results from the initial SERP were not allowed to do more HITs. As in Study 1, to avoid learning effects, the assignment of participants to experimental conditions was randomized dynamically.

4. RESULTS

Our goal was to investigate the effect of aggregated search coherence on search behavior. More specifically, we investigated four research questions: (**RQ1**) the effect of the query-senses represented in the web results on user interaction with the web results, (**RQ2**) the effect of the query-senses represented in the image results on user interaction with the web results, (**RQ3**) possible interaction effects between the query-senses represented in the web results and the image results, and (**RQ4**) the effect of the query-senses represented in the image results on user interaction with state-of-the-art web results.

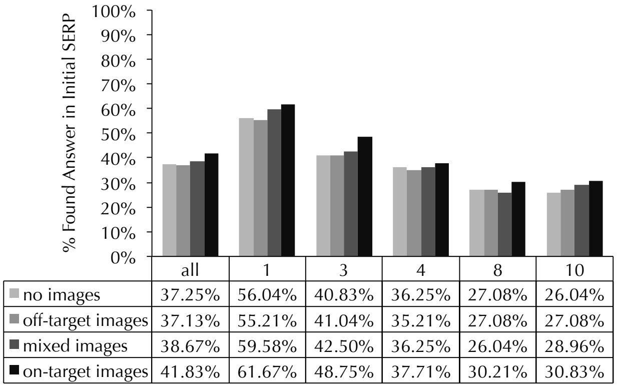
4.1 Study 1 Results

In Study 1, we manipulated the query-sense distribution associated with the image results and the web results presented in the initial SERP. Again, the image results were manipulated by presenting no image results, all image results on the target query-sense (consistent with the search task description), all image results on the off-target query-sense (inconsistent with the search task description), and an even mix of the two. The web results were manipulated by showing 9/10 off-target web results and varying the position of the answer page between ranks 1, 3, 4, 8, and 10.

User interaction with the web results was operationalized using two outcome measures:

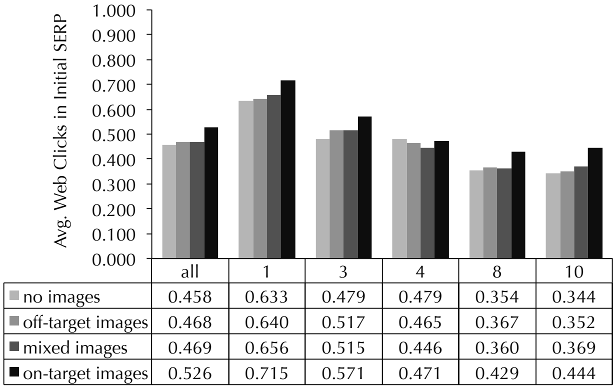
- Found answer (binary-valued): equals 1 or 0 depending on whether the participant found the answer page in the initial SERP and selected this page as containing the requested information.
- Number of web clicks (real-valued): number of times the participant clicked on a web result presented in the initial SERP. This measure included clicks on the answer page as well as clicks on other web results in the initial SERP.

Results for Study 1 are presented in Figure 4 in terms of percentage of experimental units where the participant found the answer page in the initial SERP (Figure 4(a)) and the average number of times participants clicked on a web result in the initial SERP (Figure 4(b)). Results are presented for all search sessions (all) and separately for those search sessions where the answer page was presented in rank 1 3, 4, 8, and 10. With respect to the average number of web clicks, the averages are less than one due to some participants leaving the initial SERP without clicking on any of its web results.



	all	1	3	4	8	10
■ no images	37.25%	56.04%	40.83%	36.25%	27.08%	26.04%
■ off-target images	37.13%	55.21%	41.04%	35.21%	27.08%	27.08%
■ mixed images	38.67%	59.58%	42.50%	36.25%	26.04%	28.96%
■ on-target images	41.83%	61.67%	48.75%	37.71%	30.21%	30.83%

(a) Found Answer Page in Initial SERP (binary-valued)



	all	1	3	4	8	10
■ no images	0.458	0.633	0.479	0.479	0.354	0.344
■ off-target images	0.468	0.640	0.517	0.465	0.367	0.352
■ mixed images	0.469	0.656	0.515	0.446	0.360	0.369
■ on-target images	0.526	0.715	0.571	0.471	0.429	0.444

(b) Number of Web Clicks in Initial SERP (real-valued)

Figure 4: Study 1 results in terms of (a) the percentage of search sessions where the participant found the answer page in the initial SERP and (b) the average number of web-clicks (on the initial SERP) per search session. Results are presented for all search sessions (all) and separately for those search sessions where the answer was presented at rank 1, 3, 4, 8, and 10. The total number of search sessions (all) was 9,600. The total number of search sessions for each answer rank condition was 1,600.

Figures 4(a) and 4(b) show several important trends. As might be expected, both outcome measures were greater when the answer page was presented higher in the initial SERP. In Figures 4(a) and 4(b), this trend can be observed for all image conditions. The main effect of *answer rank* on both outcome measures was statistically significant (found answer page: $\chi^2(4) = 515.11$, $p < 0.001$; number of web clicks: $F(4, 9599) = 76.18$, $p < 0.001$). Bonferroni-adjusted post-hoc comparisons were performed and, for both outcome measures, significant differences were found between all answer-rank pairs except between ranks 8 and 10. Thus, with respect to **RQ1**, Study 1 results show that the query-senses represented in the web results (and their ranks) have an effect on user interaction with the web results.

In terms of **RQ2**, both outcome measures were greater when the image results were more consistent with the search task (see "all" in Figures 4(a) and 4(b)). The main effect of *blended images* on both outcome measures was statistically significant (found answer page: $\chi^2(3) = 14.57$, $p =$

0.002; number of web clicks: $F(3, 9599) = 6.32$, $p < 0.001$). Bonferroni-adjusted post-hoc comparisons show that *found answer* was significantly higher in the on-target condition than in the no images and off-target images conditions, and that *number of web clicks* was significantly higher in the on-target condition than for *all* other image conditions.

These results show that the images can affect user interaction with the web results. More specifically, when the senses represented in the image results were more consistent with the search task, users interacted with the web results more. For example, when the images were all on-target, participants were 12% more likely to find the answer page than when the images were all off-target (a change from 37.13% to 41.83%) and 8% more likely than when the images had mixed senses (a change from 38.67% to 41.83%). Likewise, when the images were all on-target, participants clicked on 12% more web results in the initial SERP than when the images were all off-target (a change from 0.468 to 0.526) and 12% more than when the images had mixed senses (a change from 0.469 to 0.526).

The differences between showing mixed images, all off-target images, or no images at all were not significant. This may be due to the fact that 9/10 web results in the initial SERP (all except the answer page) were on the off-target sense. From our data, it appears that in order to affect users' decisions to interact with web results more, the images needed to all be on-target.

Figure 4 also shows the effect of images when the answer page was presented in different ranks. Two important results stand out. First, the images had an effect on both outcome measures even when the answer page was presented in rank 1 and in rank 3 (above the image results, if presented). For example, when the answer page was presented in rank 1, participants were 12% more likely to find it when the images were on-target vs. off-target (a change from 55.21% to 61.67%). When the answer page was presented in rank 3, participants were 19% more likely to find it when the images were on-target vs. off-target (a change from 41.04% to 48.75%). The effect of images was present even when the answer page was ranked above the images.

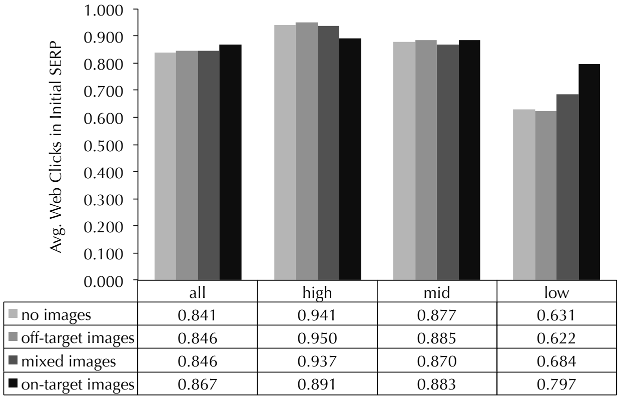
The second trend worth noting is that when the answer page was presented in rank 4, the images had less of an effect on both outcome measures. For example, when the answer page was presented in rank 4, participants were only 7% more likely to find it when the images were on-target vs. off-target (a change from 35.21% to 37.71%). This suggests that participants noticed the images irrespective of the results and subsequently made a brief downward scan, finding the answer page in rank 4. Although our statistical analysis did not show a significant interaction effect, Figure 4 shows that the image condition had less of an effect when the answer page was presented in rank 4 than in the other ranks.

To summarize, Study 1 results show a main effect from the query-senses represented in the web results (the rank of the answer page) and the query-senses represented in the image results. These results also show that the degree of influence from the image results depends on the web results. The effect of the image results is diminished when the web results contain a good result in a position that is likely to be noticed given other information shown in the results page (in our case, when the answer page was presented immediately below the visually salient images).

4.2 Study 2 Results

In Study 2, we also manually manipulated the blending of image results in the initial SERP and their query-sense distribution (no images, all off-target, mixed, and all on-target). The web results, however, were not manipulated. Instead, we simply presented the algorithmic web results returned by the Bing Web Search API (which were cached in advance). As previously described, in order to study the interaction between the images and the web results, our 80 experimental search tasks were grouped into bins based on the number of web results on the target sense and their ranks. User interaction with the web results was operationalized using the number times a participant clicked on a web result presented in the initial SERP.

Results for Study 2 are presented in Figure 5 for all search sessions combined (all) and separately for those search sessions associated with each bin (high, mid, and low). Results are shown in terms of the average number of web-result clicks (on the initial SERP) per search session. All numbers are slightly less than one due to some participants leaving the initial SERP without clicking on any web results.



	all	high	mid	low
■ no images	0.841	0.941	0.877	0.631
■ off-target images	0.846	0.950	0.885	0.622
■ mixed images	0.846	0.937	0.870	0.684
■ on-target images	0.867	0.891	0.883	0.797

Figure 5: Study 2 results in terms of the average number of web-clicks (on the initial SERP) per search session. Results are presented for all search sessions (all) and separately for those search sessions assigned to each bin. The total number of search sessions (all) was 6,400. The high and low bins had 1,280 sessions each and the mid bin had 3,840.

Several results are worth noting. With respect to **RQ1**, results show a greater number of web-result clicks (on the initial SERP) during search tasks from higher bins. In other words, user interaction with the web results in the initial SERP was higher when more of the top web results presented in the initial SERP were on the target sense (consistent with the search task description). A one-way ANOVA shows that the number of web clicks were significantly different for search sessions associated with different bins ($F(2, 6399) = 54.67$, $p < 0.001$). Bonferroni-adjusted post-hoc comparisons show significant differences between all bin-pairs: high vs. mid ($p = 0.047$), mid vs. low ($p < 0.001$), and high vs. low ($p < 0.001$).

With respect to **RQ2**, if we consider all search tasks together (all), the images had only a small effect on user interaction with the web results. A one-way ANOVA shows no significant difference in the number of web clicks (on the initial SERP) across different image conditions ($F(3, 6399) = 0.52$, $p = 0.666$). If we consider each bin separately, however, we see a different result. While the images did not have an effect on search sessions associated with the high and mid bins, they did have an effect on sessions associated with the low bin. A two-way ANOVA shows a significant interaction effect of image condition and bin-assignment on the number of web clicks ($F(6, 6388) = 2.47$, $p = 0.022$). A simple main effect analysis shows that the image condition had an effect in the low bin ($p = 0.002$), but not in the high bin ($p = 0.662$) or in the mid bin ($p = 0.954$). Furthermore, bin-assignment had a significant effect in the no images ($p < 0.001$), off-target images ($p < 0.001$), and mixed images conditions ($p < 0.001$), but no significant effect in the on-target images condition ($p = 0.096$). Bonferroni-adjusted post-hoc comparisons show that within the low bin, there were significantly more web clicks in the on-target images condition than in the off-target images ($p = 0.001$) and no images conditions ($p = 0.001$).

To summarize, Study 2 results show a main effect from the query-senses represented in the web results and their ranks, but no main effect from the query-senses represented in the image results. However, we did find an interaction effect between the web and image results. The image results had a greater effect when the web results in the initial SERP were the least consistent with the search task.

5. DISCUSSION

In Study 1, the web results in the initial SERP were manipulated by showing all off-target web results except for the answer page, which was presented either in rank 1, 3, 4, 8, and 10. One concern is whether participants became aware of this manipulation and knew where to look for the answer page. We do not believe this was case, for two reasons. First, most experimental units were completed by participants who did not complete enough experimental units to discover the manipulation. A total of 875 AMT workers participated in Study 1. About 30% of all experimental units were completed by participants who completed 15 or less, 53% were completed by participants who completed 30 or less, and 77% were completed by participants who completed 45 or less. Only about 7% of all experimental units were completed by participants who completed more than 75. Second, while clicks on the initial SERPs presented during an experimental task were biased towards ranks 1, 3, 4, 8, and 10, clicks on the initial SERPs presented during a distractor task (which presented algorithmic web results) followed the typical click distribution (Figure 6).

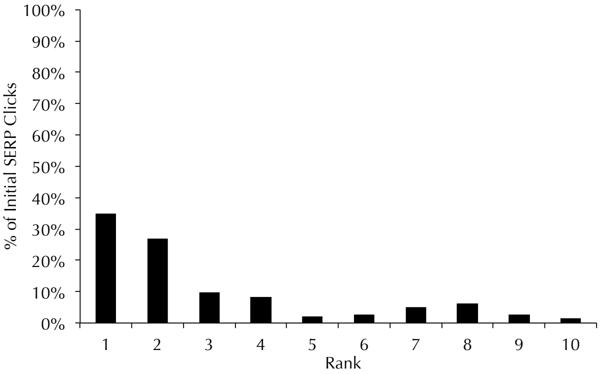


Figure 6: Click distribution on initial SERPS associated with Study 1 distractor tasks.

With respect to our original research questions (**RQ1-4**), Study 1 and Study 2 point towards the following conclusions. With respect to **RQ1**, the query-senses represented in the web results have a clear effect on user interaction with the web results. This trend was consistent in Study 1 and Study 2. In Study 1, the position of the answer page had a significant effect on both outcome measures. In study 2, the bin assignment had a significant effect on the number of web clicks in the initial SERP. With respect to **RQ2** and **RQ3**, results from Study 1 and Study 2 suggest that the image results can have a significant effect on user interaction with the web results. It depends, however, on the query-senses represented in the web results. When many of the top web results are on the target sense, such as in the experimental units assigned to the high and mid bins in Study 2, then the effect of images is low. However, when only a few of the top web results are on the target sense, such as in all experimental units in Study 1 and those assigned to the low bin in Study 2, then the effect of images is high. With respect to **RQ4**, Study 2 results show that, under certain conditions (those associated with the low bin), the images can have an effect on user interaction with state-of-the-art web results (in our case, those retrieved by the Bing Web Search API). In general, the effect of the images is likely to be greater when the web results are highly diversified, with only one or two web results consistent with the user's intended query-sense.

Our results have important implications for different areas of IR research. Given a query, search engines return various different types of information: web results, vertical results, query suggestions, advertisements, etc. Our results show that one component of the SERP can affect user interaction with a different component of the SERP. The challenge is that different types of information often originate from *independent* systems. Within aggregated search, cross-component coherence should be taken into consideration when designing new techniques for vertical selection and presentation as well as new methods for evaluation. Our results show that not all false-positive predictions are equal. While our participants were not looking for images, presenting on-target vs. off-target images increased their interaction with the web results. Finally, our results have implications for work in diversity ranking. We found that the web results that were most susceptible to the effect of images were those with only a few on-target web results in the top ranks. Thus, the degree of diversification in the web results may need to consider the degree of diversification within other components on the SERP.

Study 1 and Study 2 were designed to determine whether the image results can affect user interaction with the web results. Neither study, however, was designed to understand why this effect takes place. One explanation is that users assume that the query-sense distributions from different systems must be related (whether or not they understand that aggregated results come from different back-end systems). Users may assume that if all the images are on target (consistent with the search task), then all the web results must be on-target, and that if all images are off-target (inconsistent with the search task), then all the web results must be off-target. A second explanation is that incoherent aggregated results require a greater cognitive effort to process. Users must learn that different senses of the query exist and must reason about how to distinguish between them. Given a set of incoherent results, the path of least effort for users might be to issue a more specific query in hopes of more coherent results. Future work may investigate the root of this effect.

6. CONCLUSION

Our objective was to determine whether image results can affect user interaction with the web results. Our results indicate that images can have an effect, but that it depends on the web results. The images have a stronger effect when the web results are not heavily skewed towards the user's desired query sense.

Several factors are likely to affect the *magnitude* with which results from one particular vertical can influence user interaction with the web results (or results from other verticals). First, in this study, we focused on image results, which are visually salient. It remains to be seen whether these results generalize to text-based verticals like news. Second, we always blended the image results between web ranks 3 and 4. The placement of the vertical results is likely to have an effect. Third, the manner in which the vertical results are combined with the web results may also be a factor. Currently, the solution of choice is to blend vertical results into the web results. Other ways of combining results are possible, for example, by presenting results from different systems in clearly marked sections or blocks. A blocked interface may make it more explicit that different types of results come from independent systems and should be judged independently. Finally, attributes of the user and the search task may factor in as well. More experienced searchers may be less affected by aggregated search coherence. Or, aggregated search coherence may affect search behavior less in situations where the user is unaware of the multiple senses of the query, for example, when a user is conducting an exploratory search to learn about an unfamiliar topic. These are all open questions for future research.

7. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM 2009*, pages 5–14. ACM, 2009.

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

[3] J. Arguello, F. Diaz, J. Callan, and B. Carterette. A methodology for evaluating aggregated search results. In *ECIR 2011*, pages 141–152. Springer Berlin / Heidelberg, 2011.

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

[5] J. Arguello, F. Diaz, and J.-F. Paiement. Vertical selection in the presence of unlabeled verticals. In *SIGIR 2010*, 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 2010*, pages 105–114. ACM, 2010.

[7] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR 1998*, pages 335–336. ACM, 1998.

[8] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *CIKM 2009*, pages 621–630. ACM, 2009.

[9] C. L. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *WSDM 2011*, pages 75–84. ACM, 2011.

[10] C. L. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *ICTIR 2009*, pages 188–199. Springer-Verlag, 2009.

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

[12] M. P. Couper, R. Tourangeau, and K. Kenyon. Picture This!: Exploring Visual Effects in Web Surveys. *Public Opinion Quarterly*, 68(2):255–266, 2004.

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

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

[15] A. C. König, M. Gamon, and Q. Wu. Click-through prediction for news queries. In *SIGIR 2009*, pages 347–354. ACM, 2009.

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

[17] J. Meyers-Levy and B. Sternthal. A two-factor explanation of assimilation and contrast effects. *Journal of Marketing Research*, 30(3):359–368, 1993.

[18] A. C. Morales and G. J. Fitzsimons. Product contagion: Changing consumer evaluations through physical contact with disgusting products. *Journal of Marketing Research*, 44(2):272–283, 20070501.

[19] 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 2011*, pages 67–76. ACM, 2011.

[20] 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 2011*, pages 715–724. ACM, 2011.

[21] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML 2008*, pages 784–791. ACM, 2008.

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

[23] M. Sherif and C. Hovland. *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change.* Yale University Press, 1961.

[24] T. A. Shimp, E. W. Stuart, and R. W. Engle. A program of classical conditioning experiments testing variations in the conditioned stimulus and context. *Journal of Consumer Research*, 18(1):pp. 1–12, 1991.

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

[26] V. Toepoel and M. P. Couper. Can verbal instructions counteract visual context effects in web surveys? *Public Opinion Quarterly*, 75(1):1–18, 2011.

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

[28] X. Zhu, A. Goldberg, and D. Van Gael, Jurgenand Andrzejewski. Improving diversity in ranking using absorbing random walks. In *HLT/NAACL 2007*. ACL, 2007.