Factors Affecting Aggregated Search Coherence and Search Behavior

Jaime Arguello, Robert Capra, and Wan-Ching Wu

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

ABSTRACT

Aggregated search is the task of incorporating results from different search services, or *verticals*, into the web search results. Aggregated search *coherence* refers to the extent to which results from different sources focus on similar senses of a given query. Prior research investigated aggregated search coherence between images and web results. A user study showed that users are more likely to interact with the web results when the images are more consistent with the intended query-sense. We build upon this work and address three outstanding research questions about aggregated search coherence: (1) Does the same "spill-over" effect generalize to other verticals besides images? (2) Is the effect stronger when the vertical results include image thumbnails? and (3) What factors influence if and when a spill-over occurs from a user's perspective? We investigate these questions using a large-scale crowdsourcing study and a smaller-scale laboratory study. Results suggest that the spill-over effect occurs for some verticals (images, shopping, video), but not others (news), and that including thumbnails in the vertical results has little effect. Qualitative data from our laboratory study provides insights about participants' actions and thought-processes when faced with (in)coherent results.

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, search behavior, evaluation, mixed-methods

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright 2013 ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2505693.

1. INTRODUCTION

Commercial search services such as Google, Bing, and Yahoo! provide access to a wide range of services besides web search. These specialized services, or *verticals*, include search engines for a specific type of media (e.g., images and video) or a specific type of search task (e.g., news and shopping). The goal of *aggregated search* is to provide integrated access to all these different systems from a single search box. From a system perspective, this is often treated as a two-part task: First, predicting which verticals to present (*vertical selection*) [4, 5, 10, 13], and second, predicting where in the web results to present them (*vertical presentation*) [3, 15, 16]. Usually, a vertical is presented by blending a few of its top results somewhere above, within, or below the first page of web results. The goal is to satisfy the user with the blended vertical results, or to convey how their information need might be satisfied by searching the vertical directly.

Aggregated search *coherence* refers to the extent to which results from different sources focus on similar senses of the query. For example, suppose a user enters the ambiguous query "saturn" and the system decides to blend image vertical results into the web results. If the images are mostly about Saturn the planet and the web results are mostly about Saturn the car, then the aggregated results have a *low* level of coherence. Conversely, if both sets of results are on the same query-sense or have a similar distribution of both, then the aggregated results have a *high* level of coherence.

We investigate the effects of aggregated search coherence on search behavior. Specifically, how do the query-senses in the vertical results affect user interaction with the web results? Prior work by Arguello and Capra [2] investigated aggregated search coherence with two components: images and web results. A user study found that given an ambiguous query (e.g., "explorer"), users are more likely to interact with the web results when the blended images are consistent with the user's intended query-sense. A user looking for information about the Space Shuttle Explorer is more likely to interact with the web results when the blended images contain pictures of the Space Shuttle instead of the Ford Explorer SUV. We refer to this as a "spill-over" effect.

We report on two user studies that investigate the effects of aggregated search coherence on search behavior and address three research questions. Our first research question (RQ1) is whether the same spill-over effect observed by Arguello and Capra [2] happens for other verticals besides images. We focus on a total of four verticals: images, news, shopping, and video.

One important characteristic of aggregated search is that results from different sources are associated with different surrogate representations. For example, image results are represented using thumbnails; video results are represented using still-frames with additional textual meta-data such as the title and duration; and news results are represented using the article's title, summary, source, and publication date, and may include a thumbnail image about the news story. Prior work found that users interact more with blended verticals that have higher visual salience (e.g., video) [18]. Thus, an important research question is: How does visual salience influence the spill-over effect from the vertical to the web results? We examine this in our second research question (RQ2) by investigating whether including image thumbnails in the vertical results moderates the spill-over effect. Are the vertical results more influential when they include image thumbnails?

Finally, while prior work observed that the query-senses in the vertical results can affect user interaction with the web results, we still know little about how this happens or why. Thus, our third research question (RQ3) investigates what factors influence if and when a spill-over occurs from a user's perspective.

Our three research questions were investigated in two user studies (Study 1 and Study 2). Study 1 was run on Amazon's Mechanical Turk using a large pool of remote participants and investigated our first two research questions. To investigate the impact of image thumbnails (RQ2), two of our verticals (news and shopping) had two versions: a text-only version and a thumbnail-augmented version. In this respect, we were able to isolate the contribution of thumbnails on user interaction with the web results. Study 2 investigated RQ3, and was designed to collect qualitative data about users' actions and thought processes while interacting with our experimental system. Thus, it was run as a smaller laboratory study.

2. RELATED WORK

Current approaches to vertical selection and presentation do not explicitly consider aggregated search coherence. This is largely true in terms of the types of evidence used by existing algorithms to make predictions and in terms of the evaluation methods and metrics used to tune systems and compare their performance. State-of-the-art algorithms for vertical selection and presentation use machine learning to combine a wide range of features, including features derived from the query string [4, 10, 15, 16], from the vertical query-log [3, 4, 9, 10], from the vertical results [3, 4, 9, 10], and from user interactions associated with previous impressions of the query [15, 16]. The present study suggests that features that indicate the level of coherence with other components of the SERP may also have potential benefits.

Aggregated search coherence is also not considered in existing methods for evaluation. The goal of vertical selection is to make binary relevance predictions for a set of candidate verticals. Vertical selection algorithms are typically evaluated using vertical relevance judgements and metrics such as accuracy [4, 5, 10] or precision and recall for each vertical independently [13]. While these metrics are easy to interpret, they assume that all false positive predictions are equal. In this work, we show that this is not the case. Depending on the vertical results, a non-relevant vertical can *also* affect user interaction with other components of the SERP.

Methods for end-to-end evaluation, which includes vertical selection and presentation, fall under three categories: on-line, test-collection, and whole-page evaluation methods. On-line methods are used to evaluate systems in a production environment using click-through data as implicit feedback. If a vertical is presented, a vertical *click* indicates a true positive prediction and a vertical *skip* indicates a false positive prediction [9, 15, 16]. Evaluating based on vertical clicks and skips suffers from the same limitation mentioned above—a false positive prediction, signaled by a vertical skip, may deserve special treatment if it also affects user interaction with other components of the SERP.

Test-collection methods follow a Cranfield-style evaluation [8]. That is, they use a set of queries with (web and vertical) relevance labels and metrics that operate on an aggregated ranking of results. Zhou *et al.* [19] proposed a test-collection-based metric that accounts for three distinguishing properties between verticals: (1) the likelihood of noticing the vertical results, (2) the expected effort in assessing their relevance, and (3) the expected gain from their relevance to the task. Our experiments suggest a fourth aspect to consider: the spill-over effect from a particular set of vertical results to other components of the SERP.

Bailey *et al.* [6] proposed a whole-page evaluation methodology referred to as *Student Assignment Satisfaction Index* (SASI). The method focuses on eliciting human judgments on parts of the SERP within the context of the whole. While query-sense coherence between different components of the SERP (including vertical results) is mentioned as an important aspect of whole-page quality, its effect on user behavior and satisfaction was not investigated [6].

Low aggregated search coherence occurs when the results from different sources focus on different senses of an ambiguous query. Several trends from prior work suggest that this is a practical problem. The first trend is that users issue ambiguous queries. Sanderson [17] compared queries from a commercial query-log to ambiguous entities in Wikipedia and WordNet and found that 4% of all unique queries and 16% of all unique head queries (the most frequent) were ambiguous in terms of their intended query-sense.

The second trend relates to the aggregated search system, especially considering that current algorithms do not explicitly enforce coherence. Given an ambiguous query ("snow leopard"), one strategy for any search engine is to diversify results (to return results about the "snow leopard" the animal and the operating system). That said, within aggregated search, nothing guarantees that the top results from different sources will have the same query-sense distribution. First, different collections may have different distributions. For instance, a collection of images will probably have more images of "snow leopard" the animal, while a collection of products for sale will probably have more product pages about the operating system. Second, the ranking algorithms used by different sources often favor different types of results. For instance, a news vertical may favor articles that are recent [9], while a local vertical may favor businesses that are geographically close [1]. Thus, given an ambiguous query, even if different collections have a similar query-sense distribution, the top-ranked results from different systems may be associated with different senses—those favored by that particular ranker.

While we focus on the spill-over from vertical to web results, prior work considered the spill-over from advertisements on the SERP to the web results. Kalyanaraman and Ivory [11] found that ad relevance (the extent to which the topic of the ad matches the topic of the query) can affect users' attitudes towards the web search results. However, this effect was only present when the ad included a thumbnail image in addition to the text. This result helps motivate our second research question (RQ2), which considers whether including thumbnails in the vertical results moderates the spill-over to the web results.

3. METHOD AND MATERIALS

We report on two user studies (Study 1 and Study 2). Both studies involved participants completing a series of search tasks using a live search engine. Study 1 was aimed at answering our first two research questions (RQ1 and RQ2) and was run on Amazon's Mechanical Turk using a large pool of participants.[1] Study 2 was aimed at answering our third research question (RQ3) and, in order to gather qualitative data about participants actions and thought processes, was run as a laboratory study. In the next sections, we describe the experimental design that was common to both user studies (Section 3.1), our experimental variables (Section 3.2), and our study materials (Sections 3.3 and 3.4). Finally, we describe Study 1 and Study 2 in more detail in Sections 3.5 and 3.6.

3.1 Experimental Design

The experimental protocol used in both studies 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 band The Eagles." or "Find information about the life cycle of a beetle." The live search engine was implemented using the Bing Web Search API.

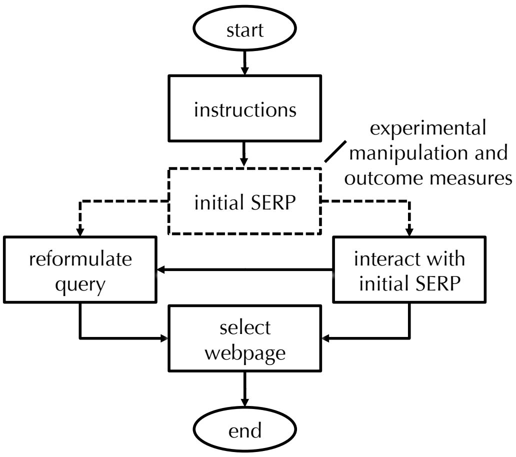

start
instructions
experimental manipulation and outcome measures
initial SERP
reformulate query
interact with initial SERP
select webpage
end

Figure 1: Experimental protocol.

Our goal was to study search behavior under the following situation: First, the user has a search task in mind ("Find information about the plot of the 2009 movie New Moon.") and issues to the system an ambiguous query ("new moon", which could mean the movie or the lunar phase). Then, in response to this query, the system returns web results and blended vertical results with a particular query-sense distribution. Finally, based on these results, the user must decide to interact with the web results or reformulate the query. In order to study this situation in a controlled environment, participants were told that, "to help you get started [with the search task], you will be provided with an initial query and a set of results". This SERP, called the *initial SERP*, is where the experimental manipulation took place.

Each search task proceeded as follows. After reading a set of instructions, participants were routed to the initial SERP, which 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 (e.g., "new moon") and the initial results included web results and, depending on the experimental condition, results from one of four verticals (images, news, shopping, and video) blended between web ranks three and four. 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. All the web and vertical results were cached in advance.

From the initial SERP, participants were instructed to search naturally by either examining the results provided or by issuing their own queries. Clicking on a result (either a result in the initial SERP or a SERP from a participant's query) opened the landing page inside an HTML frame. A button was displayed above the frame labeled: "Click here if this webpage contains the requested information". Clicking this button ended the search task. Beyond starting from the initial SERP, participants were given no other constraints. They were free to examine any result and to use the browser back button to continue searching and issuing new queries. Participant queries returned results from the Bing Web Search API without vertical results. In both studies, participants were told that our goal was to test a new search engine.

Our goal was to investigate whether users are more likely to interact with the web results when the vertical results are more consistent with the intended query-sense. Thus, in both studies, our main focus was on participants' responses to the initial SERP. In Study 1, we focused on two binary-valued outcome measures: (1) Did the participant click on a web result from the initial SERP? and (2) Did the participant ultimately select a web result from the initial SERP as containing the requested information? In Study 2, we focused on participants' think-aloud comments and interview responses about their actions on the initial SERP.

3.2 Experimental Variables

Between Studies 1 and 2, we manipulated the following experimental variables: *vertical*, *vertical query-sense*, and *vertical image*. These variables controlled which vertical was blended into the initial SERP and how it was displayed.

The *vertical* variable manipulated the vertical blended into the initial SERP. In total, we experimented with four different verticals: images, video, news, and shopping. Study 1 used all four and Study 2 focused on images and news. The images, video, and news verticals were implemented using search APIs provided by Bing and the shopping vertical was implemented using the eBay Product Search API. Screen

[1] MTurk is a crowdsourcing marketplace where requesters can publish simple tasks, called human intelligence tasks or *HITs*, for workers to complete in exchange for compensation.

shots of blended results from all four verticals are shown in Figure 2. The blended vertical results were designed to look similar to how they look in commercial systems such as Google, Bing, and Yahoo!. Image results were displayed using thumbnails; video results were displayed using still-frames and included the video's title and duration; news results were displayed using the title, summary snippet, news source, and publication age; and shopping results were displayed using the product name, price, and condition (new, used). As described in more detail below, the news and shopping verticals had two versions: a text-only version and a version that also displayed image thumbnails in addition to the text. For the news vertical, we used a reduced-sized rendering of an image pulled from the underlying webpage. For the shopping result, we used the image of the product provided by the eBay API.

The *vertical query-sense* variable manipulated the query-senses represented in the blended 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 inconsistent with the search task description, and mixed results had a combination of on-target and off-target results. For the images, shopping, and video verticals, mixed results displayed two on-target and three off-target results. For the news vertical, mixed results displayed one on-target and two off-target results.

The *vertical image* variable manipulated the use of image thumbnails in the vertical results and had two values: images and no images. In the images condition, each blended vertical result included a thumbnail in addition to the textual components, while in the no images condition, each vertical result displayed only the textual components. It seemed unlikely that images and video results would ever be displayed without thumbnails. Therefore, the vertical image variable was only manipulated for news and shopping.

3.3 Search Task Design

We constructed a total of 300 search tasks (75 search tasks per vertical). Each search task had four components: a search task description, an initial query, an on-target query-sense, and an off-target query-sense. The search task description consisted of a request for information, for example "Find information about the Proton car company." As described above, participants began each search task from an initial SERP, which included the search task description, an initial query, a set of web results, and, depending on the experimental condition, blended results from one out of four verticals. By design, the initial query (e.g., "proton") was ambiguous. The on-target query-sense was defined as the same query-sense associated with the search task description (e.g., Proton, the car company) and the off-target query-sense was defined as a different query-sense (e.g., proton, the subatomic particle).

Our search tasks were designed using a procedure similar to that used by Arguello and Capra [2] and Sanderson [17]. First, we needed to gather a set of ambiguous entities. To this end, we identified all entities associated with a Wikipedia disambiguation page. In Wikipedia, a disambiguation page serves as a navigational hub and contains links to articles about difference senses of an ambiguous entity. For instance, the disambiguation page about the entity "explorer" has links to articles on the Ford SUV, the Space Shuttle, and the web browser. A total of 122,130 Wikipedia disambiguation pages were identified using regular expressions (i.e., any Wikipedia page having "disambiguation" in the title or the "{{disambig}}" tag in the Wiki mark-up).

Second, our ultimate goal was to use ambiguous Wikipedia entities as the initial queries in our tasks. To this end, we wanted entities that a user might actually issue to a web search engine. Thus, we filtered all entities not appearing in the AOL query-log at least once. This resulted in a subset of 34,151 candidate entities.

Third, we wanted entities with a strong orientation towards one of our four verticals: images, news, shopping, and video. Four possibly overlapping subsets of entities where constructed by issuing each entity to the Bing search engine and identifying those entities that triggered the images, news, shopping, and video vertical. Of the original subset of 34,151 candidate entities, 20,080 triggered the images vertical, 7,005 triggered the news vertical, 3,814 triggered the shopping vertical, and 7,073 triggered the video vertical.

Fourth, because we planned to experimentally manipulate the query-senses in the blended vertical results, we wanted entities that might retrieve a mixture of senses in the top vertical results. Therefore, we issued each subset of entities to its corresponding API (the Bing Image, News, and Video Search API; and the eBay Product Search API) and identified 75 entities per vertical where the top results were on multiple query-senses.

Finally, for each entity, we selected one sense as the on-target sense and another as the off-target sense and constructed the search task description to be consistent with the on-target sense. Search tasks were designed to require web results rather than vertical results. Table 1 shows a few example tasks.

3.4 Vertical Images

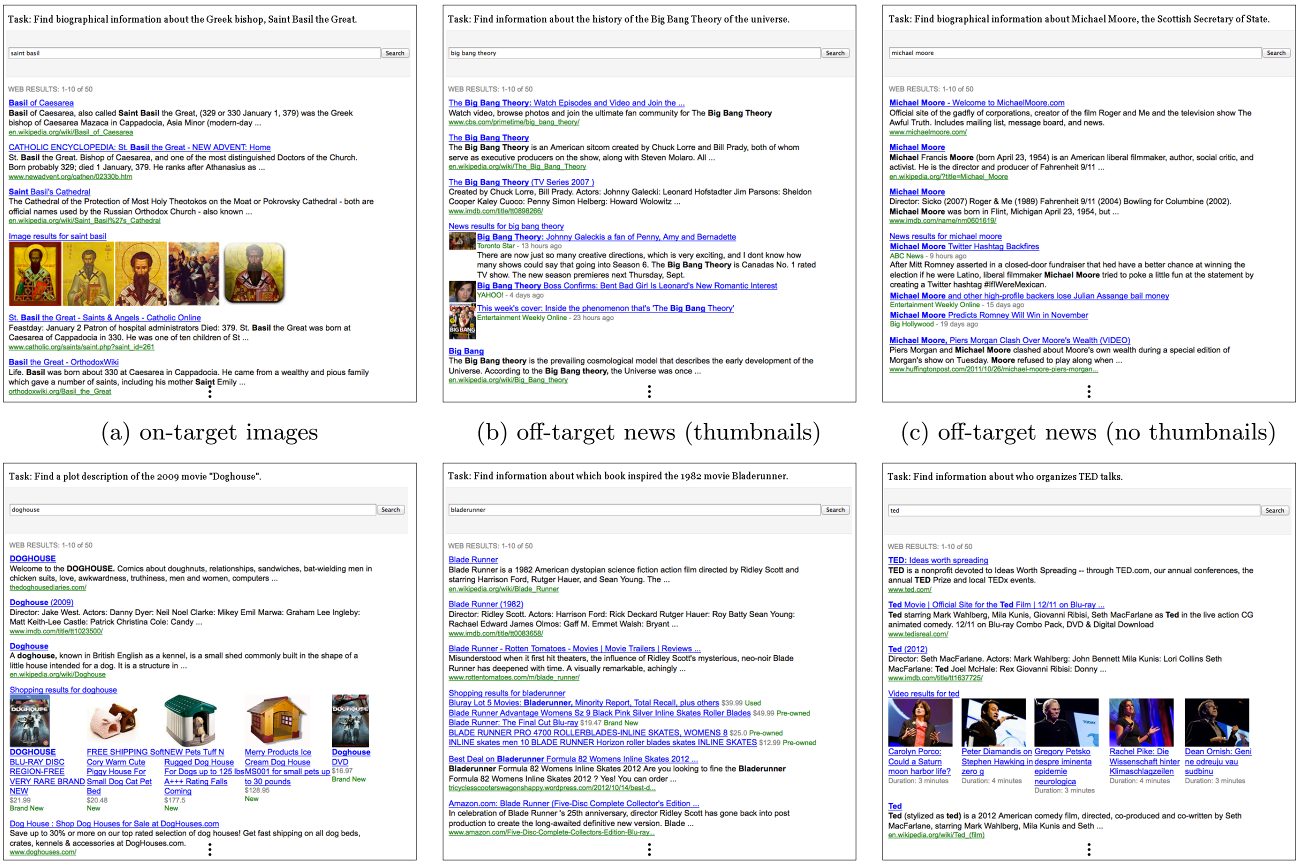
One of our goals was to investigate the impact of thumbnails on the spill-over effect (RQ2). To facilitate this analysis, two of our verticals (news and shopping) had a text-only version and a thumbnail-augmented version.

Implementing both versions of the shopping vertical was straightforward. In addition to the textual metadata associated with each result (the product title, price, and condition), the eBay Product Search API returns a thumbnail of the product. The text-only version displayed only the textual metadata and the image-augmented version displayed the thumbnails in addition to the textual meta-data.

Implementing both versions of the news vertical required more work because the Bing News Search API does not return images. To associate an image with each news result, we ran a preliminary annotation task on Amazon's Mechanical Turk. Our ultimate goal was to augment each news result with a single image pulled from the underlying webpage. The selection criterion was that the image had to strongly relate to the main topic of the news article.

The image assessment study proceeded as follows.[2] First, we cached a set of 5 on-target and 5 off-target news results for each of the 75 news tasks. Second, we used `wget` to cache every image referenced in each result. Third, for each page, a candidate set of up to ten images was selected using a rule-based classifier that scored images based on size, aspect-ratio, color distribution, file-type, and filename (many images used as icons and buttons specify this func-

[2] Due to space limitations, we cover only the necessary details of this study and the refer the reader to Capra *et al.* [7].



Task: Find biographical information about the Greek bishop, Saint Basil the Great.

saint basil | Search

WEB RESULTS: 1-10 of 50

Basil of Caesarea
Basil of Caesarea, also called **Saint Basil** the Great, (329 or 330 January 1, 379) was the Greek bishop of Caesarea Mazaca in Cappadocia, Asia Minor (modern-day ...
en.wikipedia.org/wiki/Basil_of_Caesarea

CATHOLIC ENCYCLOPEDIA: St. **Basil** the Great - NEW ADVENT: Home
St. **Basil** the Great. Bishop of Caesarea, and one of the most distinguished Doctors of the Church. Born probably 329; died 1 January, 379. He ranks after Athanasius as ...
www.newadvent.org/cathen/02330b.htm

Saint Basil's Cathedral
The Cathedral of the Protection of Most Holy Theotokos on the Moat or Pokrovsky Cathedral - both are official names used by the Russian Orthodox Church - also known ...
en.wikipedia.org/wiki/Saint_Basil%27s_Cathedral

Image results for saint basil

St. **Basil** the Great - Saints & Angels - Catholic Online
Feastday: January 2 Patron of hospital administrators Died: 379. St. **Basil** the Great was born at Caesarea of Cappadocia in 330. He was one of ten children of St ...
www.catholic.org/saints/saint.php?saint_id=261

Basil the Great - OrthodoxWiki
Life. **Basil** was born about 330 at Caesarea in Cappadocia. He came from a wealthy and pious family which gave a number of saints, including his mother **Saint** Emily ...
orthodoxwiki.org/Basil_the_Great

⋮

(a) on-target images

Task: Find information about the history of the Big Bang Theory of the universe.

big bang theory | Search

WEB RESULTS: 1-10 of 50

The **Big Bang Theory:** Watch Episodes and Video and Join the ...
Watch video, browse photos and join the ultimate fan community for The **Big Bang Theory**
www.cbs.com/primetime/big_bang_theory/

The **Big Bang Theory**
The **Big Bang Theory** is an American sitcom created by Chuck Lorre and Bill Prady, both of whom serve as executive producers on the show, along with Steven Molaro. All ...
en.wikipedia.org/wiki/The_Big_Bang_Theory

The **Big Bang Theory** (TV Series 2007)
Created by Chuck Lorre, Bill Prady. Actors: Johnny Galecki: Leonard Hofstadter Jim Parsons: Sheldon Cooper Kaley Cuoco: Penny Simon Helberg: Howard Wolowitz ...
www.imdb.com/title/tt0898266/

News results for big bang theory

Big Bang Theory: Johnny Galeckis a fan of Penny, Amy and Bernadette
Toronto Star - 13 hours ago
There are now just so many creative directions, which is very exciting, and I dont know how many shows could say that going into Season 6. The **Big Bang Theory** is Canadas No. 1 rated TV show. The new season premieres next Thursday, Sept.

Big Bang Theory Boss Confirms: Bent Bad Girl Is Leonard's New Romantic Interest
YAHOO! - 4 days ago

This week's cover: Inside the phenomenon that's 'The **Big Bang** Theory'
Entertainment Weekly Online - 23 hours ago

Big Bang
The **Big Bang theory** is the prevailing cosmological model that describes the early development of the Universe. According to the **Big Bang theory,** the Universe was once ...
en.wikipedia.org/wiki/Big_Bang_theory

⋮

(b) off-target news (thumbnails)

Task: Find biographical information about Michael Moore, the Scottish Secretary of State.

michael moore | Search

WEB RESULTS: 1-10 of 50

Michael Moore - Welcome to MichaelMoore.com
Official site of the gadfly of corporations, creator of the film Roger and Me and the television show The Awful Truth. Includes mailing list, message board, and news.
www.michaelmoore.com/

Michael Moore
Michael Francis **Moore** (born April 23, 1954) is an American liberal filmmaker, author, social critic, and activist. He is the director and producer of Fahrenheit 9/11 ...
en.wikipedia.org/?title=Michael_Moore

Michael Moore
Director: Sicko (2007) Roger & Me (1989) Fahrenheit 9/11 (2004) Bowling for Columbine (2002). **Michael Moore** was born in Flint, Michigan April 23, 1954, but ...
www.imdb.com/name/nm0601619/

News results for michael moore

Michael Moore Twitter Hashtag Backfires
ABC News - 9 hours ago
After Mitt Romney asserted in a closed-door fundraiser that hed have a better chance at winning the election if he were Latino, liberal filmmaker **Michael Moore** tried to poke a little fun at the statement by creating a Twitter hashtag #IfIWereMexican.

Michael Moore and other high-profile backers lose Julian Assange bail money
Entertainment Weekly Online - 15 days ago

Michael Moore Predicts Romney Will Win in November
Big Hollywood - 19 days ago

Michael Moore, Piers Morgan Clash Over Moore's Wealth (VIDEO)
Piers Morgan and **Michael Moore** clashed about Moore's own wealth during a special edition of Morgan's show on Tuesday. **Moore** refused to play along when ...
www.huffingtonpost.com/2011/10/26/michael-moore-piers-morgan...

⋮

(c) off-target news (no thumbnails)

Task: Find a plot description of the 2009 movie "Doghouse".

doghouse | Search

WEB RESULTS: 1-10 of 50

DOGHOUSE
Welcome to the **DOGHOUSE.** Comics about doughnuts, relationships, sandwiches, bat-wielding men in chicken suits, love, awkwardness, truthiness, men and women, computers ...
thedoghousediaries.com/

Doghouse (2009)
Director: Jake West. Actors: Danny Dyer: Neil Noel Clarke: Mikey Emil Marwa: Graham Lee Ingleby: Matt Keith-Lee Castle: Patrick Christina Cole: Candy ...
www.imdb.com/title/tt1023500/

Doghouse
A **doghouse,** known in British English as a kennel, is a small shed commonly built in the shape of a little house intended for a dog. It is a structure in ...
en.wikipedia.org/wiki/Doghouse

Shopping results for doghouse

DOGHOUSE BLU-RAY DISC REGION-FREE VERY RARE BRAND NEW	FREE SHIPPING Soft Cory Warm Cute Piggy House For Small Dog Cat Pet Bed	NEW Pets Tuff N Rugged Dog House For Dogs up to 125 lbs A+++ Rating Falls Coming	Merry Products Ice Cream Dog House MS001 for small pets up to 30 pounds	**Doghouse** DVD
$21.99	$20.48	$177.5	$128.95	$16.97
Brand New	New	New	New	Brand New

Dog House : Shop Dog Houses for Sale at DogHouses.com
Save up to 30% or more on our top rated selection of dog houses! Get fast shipping on all dog beds, crates, kennels & accessories at DogHouses.com.
www.doghouses.com/

⋮

(d) mixed shopping (thumbnails)

Task: Find information about which book inspired the 1982 movie Bladerunner.

bladerunner | Search

WEB RESULTS: 1-10 of 50

Blade Runner
Blade Runner is a 1982 American dystopian science fiction action film directed by Ridley Scott and starring Harrison Ford, Rutger Hauer, and Sean Young. The ...
en.wikipedia.org/wiki/Blade_Runner

Blade Runner (1982)
Director: Ridley Scott. Actors: Harrison Ford: Rick Deckard Rutger Hauer: Roy Batty Sean Young: Rachael Edward James Olmos: Gaff M. Emmet Walsh: Bryant ...
www.imdb.com/title/tt0083658/

Blade Runner - Rotten Tomatoes - Movies | Movie Trailers | Reviews ...
Misunderstood when it first hit theaters, the influence of Ridley Scott's mysterious, neo-noir Blade Runner has deepened with time. A visually remarkable, achingly ...
www.rottentomatoes.com/m/blade_runner/

Shopping results for bladerunner
Bluray Lot 5 Movies: **Bladerunner,** Minority Report, Total Recall, plus others $39.99 Used
Blade Runner Advantage Womens Sz 9 Black Pink Silver Inline Skates Roller Blades $49.99 Pre-owned
Blade Runner: The Final Cut Blu-ray $19.47 Brand New
BLADE RUNNER PRO 4700 ROLLERBLADES-INLINE SKATES, WOMENS 8 $25.0 Pre-owned
INLINE skates men 10 BLADE RUNNER Horizon roller blades skates INLINE SKATES $12.99 Pre-owned

Best Deal on **Bladerunner** Formula 82 Womens Inline Skates 2012 ...
Bladerunner Formula 82 Womens Inline Skates 2012 Are you looking to fine the **Bladerunner** Formula 82 Womens Inline Skates 2012 ? Yes! You can order ...
tricyclesscooterswagonshappy.wordpress.com/2012/10/14/best-d...

Amazon.com: Blade Runner (Five-Disc Complete Collector's Edition ...
In celebration of Blade Runner 's 25th anniversary, director Ridley Scott has gone back into post production to create the long-awaited definitive new version. Blade ...
www.amazon.com/Five-Disc-Complete-Collectors-Edition-Blu-ray...

⋮

(e) mixed shopping (no thumbnails)

Task: Find information about who organizes TED talks.

ted | Search

WEB RESULTS: 1-10 of 50

TED: Ideas worth spreading
TED is a nonprofit devoted to Ideas Worth Spreading -- through TED.com, our annual conferences, the annual **TED** Prize and local TEDx events.
www.ted.com/

Ted Movie | Official Site for the **Ted** Film | 12/11 on Blu-ray ...
Ted starring Mark Wahlberg, Mila Kunis, Giovanni Ribisi, Seth MacFarlane as **Ted** in the live action CG animated comedy. 12/11 on Blu-ray Combo Pack, DVD & Digital Download
www.tedisreal.com/

Ted (2012)
Director: Seth MacFarlane. Actors: Mark Wahlberg: John Bennett Mila Kunis: Lori Collins Seth MacFarlane: **Ted** Joel McHale: Rex Giovanni Ribisi: Donny ...
www.imdb.com/title/tt1637725/

Video results for ted

Carolyn Porco: Could a Saturn moon harbor life?	Peter Diamandis on Stephen Hawking in zero g	Gregory Petsko despre iminenta epidemie neurologica	Rachel Pike: Die Wissenschaft hinter Klimaschlagzeilen	Dean Ornish: Geni ne odreuju vau sudbinu
Duration: 3 minutes	Duration: 4 minutes	Duration: 3 minutes	Duration: 4 minutes	Duration: 3 minutes

Ted
Ted (stylized as **ted)** is a 2012 American comedy film, directed, co-produced and co-written by Seth MacFarlane, starring Mark Wahlberg, Mila Kunis and Seth ...
en.wikipedia.org/wiki/Ted_(film)

⋮

(f) on-target video

Figure 2: Initial SERP screenshots associated with different experimental conditions (cropped)

tion in the filename). Finally, we designed a Mechanical Turk task where participants were shown a news article and its candidate images side-by-side and were asked to "rate how well each image is related to the main topic of this webpage." Responses were indicated on a 7-point scale using radio buttons anchored with the labels *very unrelated* on the left and *very related* on the right. For each news result, we collected image annotations from five redundant workers. The HIT was priced at $0.10 USD. After the annotation effort, we associated a single image with each news result by randomly selecting from those images with an average rating of five or greater. News results without a "good" image were discarded. In the end, we were left with three on-target and three off-target news results with an associated image.

3.5 Study 1

Study 1 was aimed at answering our first two research questions: (RQ1) Does the spill-over effect generalize across different verticals (images, news, shopping, and video)? and (RQ2) Does the presence of thumbnails in the vertical results moderate the spill-over effect? Study 1 was run on Amazon's Mechanical Turk using a large pool of participants. Each HIT consisted of a single search task and, as described in Section 3.1, each search task originated from an initial SERP, which was experimentally manipulated.

Study 1 manipulated all three experimental variables described in Section 3.2: vertical, vertical query-sense, and vertical image. This resulted in a total of 1,650 experimental conditions, counted as follows. The images and video verticals always displayed thumbnails and were therefore each associated with 225 experimental conditions (75 search tasks × 3 vertical query-senses × 1 vertical image condition = 225). On the other hand, the news and shopping verticals had two versions (a text-only and thumbnail-augmented version) and were therefore each associated with 450 experimental conditions (75 search tasks × 3 vertical query-senses × 2 vertical image conditions = 450). Finally, for each of our 300 search tasks, we also collected interaction data with the initial SERP displaying only web results, for a total of 1,650 conditions (225 + 225 + 450 + 450 + 300 = 1,650). For each experimental condition, we collected interaction data from 10 redundant participants for a total of 16,500 HITs. Each HIT was priced at $0.10 USD.

While the experiment was run on MTurk, our HITs were implemented as *external* HITs, meaning that everything besides recruitment and compensation was managed by our own server. Managing our HITs locally allowed us to control the assignment of participants to experimental conditions, to detect and reject careless workers "on the fly", and to record the necessary user interactions with the initial SERP and the live search engine. Participants were assigned to experimental conditions randomly, except that the same participant could not see the same search task more than once (for different vertical query-sense and vertical image conditions).

Table 1: **Example search tasks. The search task description corresponds to the information request given to participants, 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.**

		on-target	**off-target**	
vertical	**initial query**	**sense**	**sense**	**task description**
images*	china lake	naval station	lake	Find historical information about the naval station “China Lake”.
images*	saint basil	greek bishop	cathedral	Find biographical information about the Greek bishop, Saint Basil the Great.
images*	salton	lake	appliance	Find information about the origins of Salton Sea Lake in California.
images*	wega	coffee maker	sony tv	Find information about the coffee maker company “WEGA”.
news*	durango	suv	x-box console	Find consumer reviews about the 2012 Dodge Durango.
news*	big bang theory	scientific theory	tv show	Find information about the history of the Big Bang Theory of the universe.
news*	sin city	las vegas	comic/movie	Find tourism information about Sin City (Las Vegas, Nevada).
news*	world series	baseball	poker	Find historical information about the first World Series in baseball.
shopping	torch	phone	led flashlight	Find consumer reviews for the Blackberry Torch phone.
shopping	theorema	watchmaker	cologne	Find information about the company that makes Theorema watches.
shopping	happy feet	movie	shoes	Find information about who did the voices in the movie “Happy Feet”.
shopping	sunfire	car	sub-woofer	Find information about why the Pontiac Sunfire was discontinued.
video	bladerunner	althete	movie	Find biographical information about athlete nicknamed “Blade Runner”.
video	ted	non-profit	movie	Find information about who organizes TED talks.
video	beautiful people	song	tv show	Find the lyrics to the song “Beautiful People” by Chris Brown.
video	desperado	song	movie	Find information about the meaning of the Eagles song ”Desperado”.

Quality control was done in three ways. First, we restricted our HITs to workers with at least a 95% acceptance rate and, to help ensure English proficiency, to workers within the US. Second, prior to the experiment, one of the authors judged the relevance of every web result presented in an initial SERP, and participants who marked three (initial SERP) non-relevant results as containing the requested information were not allowed to do more HITs. Finally, to obtain behavioral data from a large pool of participants, workers were not allowed to do more than 25 of our HITs. In the end, we collected data from 993 participants.

3.6 Study 2

Study 2 focused on our third result question: (RQ3) What factors influence if and when a spill-over occurs from a user's perspective? In other words, in which situations do the query-senses in the vertical results affect user engagement with the web results? In order to gather qualitative data about participants' actions in the initial SERP, Study 2 was run as a laboratory study.

Study 2 manipulated two of the variables discussed in Section 3.2: vertical and vertical query-sense. Based on results from Study 1, we focused on the two verticals that elicited the most divergent behavior from participants (images and news) and the two most extreme vertical query-sense conditions (off-target and on-target). For each vertical, we selected four tasks: two that exhibited a strong spill-over effect in Study 1 (the vertical results had a strong effect on user interaction with the web results) and two tasks that exhibited a weak spill-over effect. In total, Study 2 was associated with 16 experimental conditions (2 verticals × 4 tasks per vertical × 2 vertical query-senses × 1 vertical image condition = 16). Different from Study 1, we did not manipulate the vertical image variable. Thus, news results always included thumbnails. The search tasks used in Study 2 are marked with a '*' in Table 1.

Sixteen participants (6 male and 10 female) were recruited from our university. Participants were graduate students from various fields of study. Similar to Study 1, participants were asked to complete a series of search tasks and each search task originated from an initial SERP. Once again, participants were told that the initial SERP was provided to “help you get started” and were instructed to search naturally by examining the results provided or doing their own searches. As in Study 1, our focus was on participants' responses to the initial SERP. Participants were compensated with $20 USD.

Study 2 participants completed a series of five search tasks. The first task was a practice task used the familiarize participants with the protocol. The remaining four experimental tasks were evenly divided between images vs. news and off-target vs. on-target. Search tasks were rotated to control for possible learning and fatigue effects, and, of course, no participant did the same search task more than once.

In order to collect qualitative data about participants' actions in the initial SERP, search sessions were recorded using screen recording software and participants were asked to describe their searches using a two-stage retrospective stimulated recall method. After completing each task, we played the recorded search and asked participants to explain their actions and thought processes. At this first stage, we only prompted participants to explain what they were thinking and the only follow-up prompt was a request to “tell us more”. Our intention was to capture participants' thought processes immediately after searching and to avoid influencing their subsequent behavior by asking potentially biasing questions.

Then, after completing all five search tasks, we played each search for the participant a second time, pausing the video when necessary. At this second stage, participants were asked specific questions about their experience with the initial SERP. Our ultimate goal was to determine: (1) whether the participant noticed the blended vertical results; (2) whether they recognized the query-senses in the vertical results; and (3) whether the vertical results influenced their actions on the initial SERP and, if so, how. In order to avoid biasing their responses, we did not ask these questions directly. Instead, we asked participants to identify (in order) each result they remembered seeing on the initial SERP. Then, for each result, we asked whether they had tried to figure out what the result was about and, if not, why. Then, we asked what motivated their next action. In cases where they quickly reformulated the initial query, we asked why they thought they would not find the requested information on the page. Conversely, in cases where they examined the initial SERP more closely, we asked why they thought they would find the requested information on the page. Again, we note that this process was carefully designed to elicit answers to the underlying questions of interest, while not revealing information that would bias participants responses on subsequent tasks. All responses from participants were recorded using a second screen recording system that was running throughout the entire experiment. Finally, participant responses were analyzed as described in Section 5.

4. STUDY 1 RESULTS

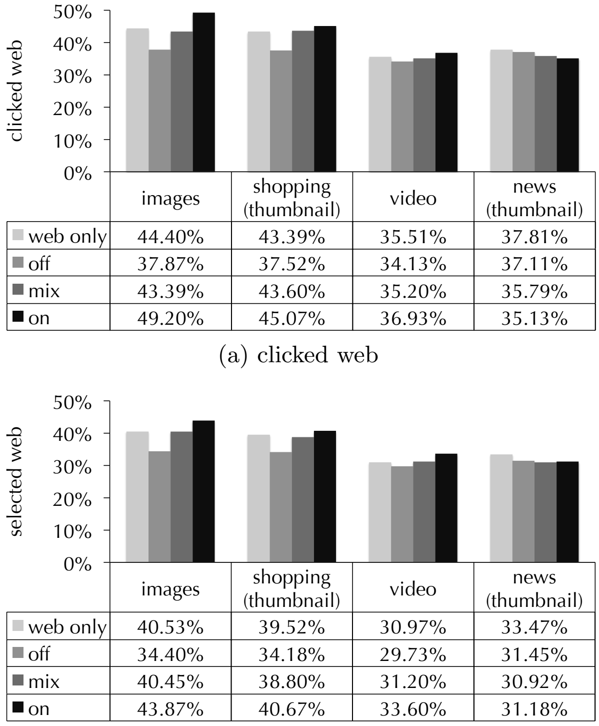
Sections 4.1 and 4.2 present our results with respect to RQ1 and RQ2, respectively.

4.1 Vertical Analysis

In this section, we examine whether the spill-over effect generalizes across verticals: images, news, shopping, and video. We focus on the versions of the news and shopping verticals that included thumbnails. User engagement with the initial SERP web results was operationalized using two binary outcome measures: (1) Did the participant click on at least one initial SERP web result? (*clicked web*) and (2) Did the participant ultimately select an initial SERP web result as containing the requested information? (*selected web*). Figures 3(a) and 3(b) show the percentage of participants for which each outcome measure was true. Recall that each vertical had its own unique set of 75 search tasks. In addition to presenting results for the on-target (on), mixed (mixed), and off-target (off) vertical query-sense conditions, we present results for when no vertical was displayed (web only). Because each experimental condition was experienced by 10 redundant participants, all percentages reported in Figure 3 are for 750 participants ($75 \times 10 = 750$).

The strongest spill-over effect was observed for the images vertical. A chi-squared test shows a significant main effect on both outcome measures (clicked web: $\chi^2(3) = 19.603$, $p < .001$; selected web: $\chi^2(3) = 14.605$, $p < .05$). For *clicked web*, post-hoc comparisons[3] show significant differences between on-target images and both mixed images ($p < .05$) and off-target images ($p < .001$), and marginally significant differences between mixed and off-target images ($p = .059$). Participants were 30% more likely to click an initial SERP web result when the images were on-target vs. off-target (an increase from 37.87% to 49.20%) and 13% more likely when the images were on-target vs. mixed (an increase from 43.39% to 49.20%). Furthermore, post-hoc comparisons show a significant difference between off-target images and the web only condition ($p < .05$). Participants were 15% less likely to click an initial SERP web result when the images were off-target than when no vertical was presented (a reduction from 44.40% to 37.87%). Showing off-target images significantly reduced user interaction with the web results compared to not showing images at all. Similar trends were observed for *selected web*.

For the shopping vertical, a chi-squared test shows a significant main effect on both outcome measures (clicked web: $\chi^2(3) = 9.856$, $p < .05$; selected web: $\chi^2(3) = 7.712$, $p = .05$). For *clicked web*, post-hoc comparisons show significant differences between on- and off-target shopping results ($p < .05$) and between mixed and off-target shopping results ($p < .05$). Furthermore, post-hoc comparisons show a significant difference between off-target shopping results and the web only condition ($p < .05$). As with the images vertical, showing off-target shopping results significantly reduced user interaction with the web results compared to not showing the vertical at all. Similar trends were observed for the *selected web* measure. However, post-hoc comparisons only showed a significant difference between on- and off-target shopping results ($p < .05$) and a marginally significant difference between off-target and web only ($p = .06$).



	images	shopping (thumbnail)	video	news (thumbnail)
▨ web only	44.40%	43.39%	35.51%	37.81%
▨ off	37.87%	37.52%	34.13%	37.11%
▨ mix	43.39%	43.60%	35.20%	35.79%
■ on	49.20%	45.07%	36.93%	35.13%

(a) clicked web

	images	shopping (thumbnail)	video	news (thumbnail)
▨ web only	40.53%	39.52%	30.97%	33.47%
▨ off	34.40%	34.18%	29.73%	31.45%
▨ mix	40.45%	38.80%	31.20%	30.92%
■ on	43.87%	40.67%	33.60%	31.18%

(b) selected web

Figure 3: Percentage of participants who (a) clicked on at least one initial SERP web result and (b) selected an initial SERP web result as containing the requested information.

For the video vertical, both outcome measures were also greater when the video results were on-target. However, while the upward trend was consistent with the images and shopping verticals, it was less pronounced. The effect of the query-sense in the video results was not significant for either measure (clicked web: $\chi^2(3) = 1.311$, $p = .726$; selected web: $\chi^2(3) = 2.73$, $p = .435$).

Finally, the trend for the news vertical was quite different. Instead of the up-ward trend observed for images, shopping, and video, the query-senses in the news results had almost no effect on either outcome measure (clicked web: $\chi^2(3) = 3.808$, $p = .283$; selected web: $\chi^2(3) = 3.072$, $p = .381$).

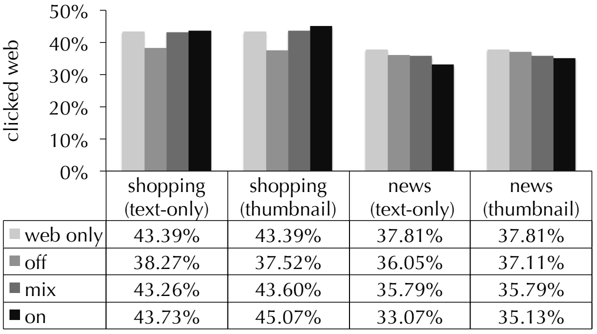
To summarize, our results show a strong and significant spill-over for the images vertical and shopping vertical; a noticeable, but non-significant spill-over for the video vertical; and no spill-over for the news vertical. We discuss possible explanations for these results in Section 6.

[3] All post-hoc comparisons were done using the modified Bonferroni correction outlined in Keppel (p.169) [12].

4.2 Thumbnail Analysis

In this section, we examine whether including thumbnail images in the shopping and news results moderates the spill-over from the vertical to the web results. Similar to the previous analysis, both outcome variables were highly correlated, so we focus our discussion on *clicked web*.

Figure 4 shows the percentage of participants who clicked on at least one initial SERP web result. We show results for shopping and news (with and without thumbnails) separately and include results for the on-target, mixed, and off-target vertical query-sense conditions as well as the web only condition. Again, the percentages are with respect to the 750 participants exposed to each condition.



	shopping (text-only)	shopping (thumbnail)	news (text-only)	news (thumbnail)
■ web only	43.39%	43.39%	37.81%	37.81%
■ off	38.27%	37.52%	36.05%	37.11%
■ mix	43.26%	43.60%	35.79%	35.79%
■ on	43.73%	45.07%	33.07%	35.13%

Figure 4: Percentage of participants who clicked on at least one initial SERP web result for the shopping and news verticals with and without thumbnails.

For each vertical, we used a logistic regression to test the interaction between vertical query-sense (on-target, mixed, off-target) and vertical image (text-only, thumbnails). The web only condition was used as the control variable.

For the shopping vertical, the *only* significant predictor was the indicator variable for the off-target vertical query-sense (Wald's $\chi^2(1) = 4.069$, $p < .05$). The odds ratio for this variable was 0.809 (less than one), which is consistent with the results shown in Figure 4. Off-target shopping results (with or without images) significantly reduced *clicked web* compared to the web only condition. The fact that no other predictor was significant suggests that thumbnails did not have a significant main effect and did not moderate the spill-over from the vertical to the web results.

For the news vertical, the logistic regression analysis showed that none of the input variables were significant predictors of *clicked web*. In other words, neither the vertical query-sense, the presence of thumbnails, nor their combination had a significant effect on *clicked web*. We discuss possible explanations for these results in Section 6.

5. STUDY 2 RESULTS

Study 2 investigated our third research question: What factors influence if and when a spill-over occurs from a user's perspective? The goal was to collect qualitative data about participants' actions and thought processes on the initial SERP. The data collected from Study 2 was analyzed as follows. Recall that Study 2 had 64 experimental search tasks (excluding the practice tasks). We reviewed the think-aloud comments and interview responses from participants and assigned each search task to one of four outcomes.

Positive and Negative spill-over: the on-target(off-target) vertical results *increased*(*decreased*) the participant's engagement with the initial SERP web results.

No spill-over: the vertical results had no effect on the participant's engagement with the initial SERP web results.

Unknown spill-over: the effect of the vertical could not be determined from the participant's comments/responses.

Of the 64 search tasks, 25 (39%) had a spill-over, 33 (52%) had no spill-over, and 6 (9%) had unknown spill-over. Of the 25 that had a spill-over, 15 had a positive and 10 had a negative spill-over. Consistent with Study 1, most spill-overs happened for the images (17) vs. the news vertical (8).

5.1 Positive Spill-over

Positive spill-overs happened in two situations. In the first, the on-target vertical results increased the participant's confidence in the overall results (including the web results).

P13: *"I saw pictures of the lake* [on-target images], *and it was an indication that I was getting close."*

P3: *"The first thing I noticed were the coffee makers* [on-target images] *...then, I clicked on the second result because I assumed it was about the coffee maker company. I didn't even read it."*

P6: *"I thought the links on the page* [web results] *would be more specific because all the images were about Saint Basil* [on-target images]."

In the second situation, we observed a more subtle process. Two participants reported lacking prior knowledge about the task and using the images to get ideas about things to search for on the SERP. Visual cues provided by the on-target images increased the participant's engagement with the web results.

P13: *"I saw the picture of the lake and the map of California* [on-target images] *...and I thought that Salton Sea must be a lake in California... But, I didn't see the California Parks and Recreation page, so I tightened-up the query."*

P11: *"When I read* [the task description], *I knew that I didn't know anything about the topic... First I saw the pictures* [on-target images], *and I saw that China Lake was a military base in the West Coast. Then I went back up and started going down the page."*

5.2 Negative Spill-over

Negative spill-overs happened exclusively because the off-target vertical results decreased the participant's confidence in the overall results (including the web results).

P7: *"I knew that these pictures were about the TV show* [off-target news thumbnails]. *That was the biggest indicator that the results were about the TV show."*

P10: *"I noticed the TVs* [off-target images] *...and knew that I should probably put something about coffee in the query."*

P10: *"I noticed the pictures* [off-target news thumbnails] *and was wondering why poker came up. Then I thought, of course 'World Series of Poker', and that's when I thought there was probably a better query."*

> P13: *"I noticed the colorful pictures* [off-target images]... *and I knew that if I clicked on a picture or something below it* [web results], *I would probably be taken to a link about the cathedral and not the saint."*

5.3 No Spill-over

Based on our analysis, there were four situations that resulted in no spill-over. In the first situation, the top web results caused the participant to not even perceive the vertical results. In these cases, participants reported immediately noticing and clicking a top-ranked *relevant* web result or immediately noticing a top-ranked *non-relevant*, off-target web result and reformulating the query.

> P15: *"I think I just looked at the top result. I looked at 'Saint Basil'* [bold text in top result summary]...*I don't remember seeing these pictures* [off-target images]."

> P14: *"I didn't look at these* [news vertical results]. *I looked at the top result and noticed that it said something about vacations* [web result about the city of Durango] *and decided that the query was too general. I didn't look at much else."*

In the second situation, the participant perceived the vertical results, but recognized them as being a specific *type* of result, not be relevant to the task. Consequently, the participant did not process the vertical results enough to notice their on-target or off-target query-sense.

> P2: *"I noticed the pictures first* [off-target images]... *But, I didn't try to figure out what they were about because I knew they wouldn't give me historical information."*

> P11 *"I looked at the news articles with the pictures* [on-target news results], *but I wanted tourism information and I know that a news article wouldn't have that.... I just discarded those."*

> P12: *"I did notice the news results again* [off-target news results], *but because I was looking for historical information* [about the first baseball World Series], *I didn't try to figure out what they were about. Now I see they're about poker."*

Third, one participant perceived the vertical results (in this case, off-target images) and processed their content enough to recognize their query-sense. However, the participant did not perceive the off-target images as being off-target.

> P14: *"I noticed these briefly* [off-target images], *and I figured this place* [Saint Basil's Cathedral] *had to be somehow related to Saint Basil."*

Finally, some participants reported noticing the vertical results, recognizing their query-sense, and trying to assess their relevance to the task. Interestingly, this happened mostly for on-target news results. This category is distinct from the second category mentioned above. In this category, we believe that the participant expended effort assessing the relevance of the on-target vertical results.

> P10: *"I looked at the images* [on-target news thumbnails] *...but nothing seemed tourism-related...that's when I just said: Forget it, I'm just gonna do a totally different query."*

> P9: *"Clearly, I was getting some Las Vegas stuff* [on-target news results], *but I thought I would need a more specific query to answer the question."*

6. DISCUSSION

With respect to RQ1, results from Study 1 suggest that the spill-over effect occurs for some verticals (images, shopping, and, to a lesser extent, video), but not others (news). While the trend for video was not significant, it was consistent with images and shopping. Insights gained from Study 2 provide possible explanations for these results. We elaborate on these at the end of this section.

With respect to RQ2, results from Study 1 also found that including thumbnails in the news and shopping results had little effect. That is, the shopping results caused a similar amount of spill-over with or without thumbnails, and the news results caused no spill-over with or without thumbnails. For shopping, including thumbnails had a small impact. As shown in Figure 4, for the on-target condition, including thumbnails increased *clicked web* from 43.73% to 45.07%, and, for the off-target condition, including thumbnails decreased *clicked web* from 38.27% to 37.81%. However, as shown in Section 4.2, this interaction was not significant.

There are at least two possible explanations for why thumbnails were not more influential. First, the vertical results were always blended between web results three and four, probably always above the fold. Thumbnails might have a greater effect when the vertical is blended below the fold, where the thumbnails might draw more attention than the text-only surrogates. Second, the amount of text associated with our news and shopping surrogates was significant. It is also possible that the textual components conveyed enough information to make the query-sense(s) recognizable. Prior work on image-augmented surrogates found that users mostly rely on the images when the textual components are poor [14].

With respect RQ3, Study 2 found different processes at play. In cases of a spill-over, participant responses suggest that the on-/off-target vertical results increased/decreased their confidence in the SERP as a whole (including the web results). During negative spill-overs, for example, participants reported noticing the off-target vertical results and thinking that the query was too broad. This is an important result. Implicit in this behavior is the *assumption* that the vertical and web results are related or coherent. These participants used the vertical results as proxy for judging the whole SERP (including the web results). Even in cases where the on-target images gave participants ideas about things to search for on the SERP, there is the implicit assumption that the vertical and the web results contain similar information. That is, the visual cues provided by the images are applicable to the web results.

Study 2 also provided examples of no spill-over. As one might expect, *at least* two things must happen for a spill-over to happen: (1) the user must perceive the vertical results and (2) must recognize their query-sense(s). When participants did not perceive the vertical, it was mostly due to the top web results—either a top web result was highly relevant (resulting in an immediate click) or highly off-target (resulting in an immediate reformulation). Arguello and Capra [2] found that images have a stronger spill-over when the web results are off-target. Our results are consistent with theirs, but suggest that *highly* off-target web results can also cause the user to reformulate before perceiving the vertical results. In cases where participants perceived, but *did not* process the vertical results, it was mostly because they recognized the vertical results as being a specific *type* of result, not relevant to the task. The participant did not process the vertical results enough to recognize the on-/off-target sense.

Finally, Study 2 also had cases where the participant perceived the vertical results, recognized their query-sense, but no spill-over happened. In one particular case, the participant did not perceive the off-target vertical results as being off-target, possibly because the on-/off-target query-senses were closely related. In most cases, and mostly for on-target news, the participant assessed the on-target vertical results, determined they were not relevant to the task, and left the initial SERP. This suggests that a positive spill-over may not occur if the user expends fruitless effort in assessing the on-target vertical results.

Results from Study 2 provide two possible explanations for why video had only a weak spill-over and news had no spill-over. At the core is the hypothesis that processing news and video results required more cognitive effort. News results were displayed using more text than the others and video results were displayed using still-frames, which do not always easily convey the main topic of the video (at least compared to how a thumbnail describes an image or a product). Our first explanation is that participants often perceived the news and video results, but did not recognize their query-senses because they required too much cognitive effort to process. All verticals were probably perceived by many participants as a specific *type* of result, not relevant to the task. However, because the news and video results required more effort to process, they were probably skipped more often than the images and shopping results. Our second explanation (not mutually exclusive with the first), is that the participants who *did* process the news and video results expended enough cognitive effort that they were soon-after ready to explore a new query. Future work might consider whether spill-overs happen more for vertical results that require less cognitive effort to process.

7. CONCLUSION

We reported on two user studies that show that the query-senses in the vertical results can affect user interaction with the web results. Results from our first study show that the "spill-over" effect happens for some verticals, but not others, and that including thumbnails in the vertical results has little effect (at least when the vertical is ranked high). Results from our second study show examples of how/why a spill-over does or does not happen from a user's perspective.

Our results suggest that a spill-over happens when users: (1) perceive the vertical, (2) recognize their on-/off-target results, and (3) allow these to increase/decrease their confidence on the SERP as a whole. Our analysis points to several factors that may influence whether a spill-over happens. The vertical rank, the vertical presentation, and the web results themselves are likely to affect whether a user even perceives the vertical. Factors such as a user's view of the vertical as a "non-web component" and the cognitive effort required in processing the vertical results may influence whether a user recognizes the vertical query-senses. And, finally, factors such as a user's topic familiarity (their ability to recognize an off-target sense), their mental model of the system, and the interface design may affect whether results from one component are used to evaluate another.

Further understanding aggregated search coherence and search behavior has important implications for aggregated search and open questions remain. Additional user studies are needed to understand which factors (of the user, the vertical, the web results, and the overall layout) determine whether a spill-over effect happens. After gaining more insight into these factors, future work should focus on designing evaluation methodologies and aggregated search algorithms that model cross-component effects.

8. REFERENCES

[1] T. Abou-Assaleh and W. Gao. Geographic ranking for a local search engine. In *SIGIR 2007*, pages 911–911. ACM, 2007.

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

[3] J. Arguello, F. Diaz, and J. Callan. Learning to aggregate vertical results into web search results. In *CIKM 2011*, pages 201–210. ACM, 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] R. Capra, J. Arguello, and F. Scholer. Augmenting web result surrogates with images. In *CIKM 2013*. ACM, 2013.

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

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

[10] 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.

[11] 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.

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

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

[14] F. Loumakis, S. Stumpf, and D. Grayson. This image smells good: effects of image information scent in search engine results pages. In *CIKM 2011*, pages 475–484. ACM, 2011.

[15] 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.

[16] 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.

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

[18] 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.

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