Augmenting Web Search Surrogates With Images

Robert Capra
School of Information & Library Science
University of North Carolina
Chapel Hill, NC, USA
rcapra@unc.edu

Jaime Arguello
School of Information & Library Science
University of North Carolina
Chapel Hill, NC, USA
jarguello@unc.edu

Falk Scholer*
School of Computer Science & Information Technology
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

ABSTRACT

While images are commonly used in search result presentation for vertical domains such as shopping and news, web search results surrogates remain primarily text-based. In this paper, we present results of two large-scale user studies to examine the effects of augmenting text-based surrogates with images extracted from the underlying webpage. We evaluate effectiveness and efficiency at both the individual surrogate level and at the results page level. Additionally, we investigate the influence of two factors: the goodness of the image in terms of representing the underlying page content, and the diversity of the results on a results page. Our results show that at the individual surrogate level, good images provide only a small benefit in judgment accuracy versus text-only surrogates, with a slight increase in judgment time. At the results page level, surrogates with good images had similar effectiveness and efficiency compared to the text-only condition. However, in situations where the results page items had diverse senses, surrogates with images had higher click precision versus text-only ones. Results of these studies show tradeoffs in the use of images in web search surrogates, and highlight particular situations where they can provide benefits.

Categories and Subject Descriptors

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

General Terms

Performance, Experimentation, Human Factors

Keywords

Search result representation, image surrogates, query ambiguity, user study, evaluation, search behavior

*Work done while the author was a visiting researcher at the School of Information and Library Science at the University of North Carolina in Chapel Hill.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505714.

1. INTRODUCTION

Search engine results are typically presented as a ranked list of surrogates. The goal of the surrogate representation is to convey the appropriate information for users to make effective and efficient relevance decisions from the search engine results page (SERP) itself, without having to examine the underlying document. For web search results, surrogates typically consist of textual elements—a title, URL, and summary snippet—from the underlying page.

Intuitively, images would appear to offer several benefits for web result surrogates. Images compactly summarize large amounts of information, provide context for understanding results, and can be quickly scanned by users. Images are commonly used in surrogates for verticals such as news and shopping. However, surrogates for web search remain primarily textual.

Prior research has explored several approaches to incorporate visual components into web result surrogates, including the use of thumbnail images [3, 13], collages of visual and textual elements [23, 25], and single images extracted from the page [12, 15, 18]. Studies of these approaches have produced mixed results, with some research showing benefits of images, and other studies finding no clear improvements over traditional text-based surrogates.

In this paper, we focus on a particular approach that has received recent attention and shown mixed results: augmenting surrogates with an image pulled from the underlying page. Prior research has suggested that this approach can produce surrogates that are favored over other methods [12] and that can achieve significant gains in effectiveness and efficiency [15]. However, other studies suggest that images play a smaller role than text in the effectiveness of the surrogate [16] and that image-augmented surrogates do not significantly improve performance versus text-only ones [18].

Our goal was to explore the effects of this approach in more depth than prior studies, and also to investigate dimensions that might influence when images help and when they hurt. We focus on two ways that images might help. First, images may convey additional information that is missing from the textual components of the surrogate. This would benefit users when making a relevance decision about the individual result represented by the surrogate. Second, images may help users during the initial process of scanning the SERP as a whole. This triage process is a time when the user is trying not only to locate a relevant result, but also to determine if the query was effective. Images could help users make these SERP-level judgments more quickly and effectively.

We also investigate two factors that may influence the effects of image-augmented surrogates. First, we consider the "goodness" of the image in representing the underlying result page. Images extracted from a webpage may or may not accurately reflect its content. Prior work has shown that image classifiers can be trained to select salient images with good accuracy [15]. However, not all pages have good images [12], and even good classifiers will make mistakes. This motivated us to investigate three points along a continuum: all good images, all bad images, and a mixture of both. We wanted to see not only if good images help, but to what extent bad images might hurt. Second, we consider how the diversity of the results on the SERP may influence the benefits of adding images to surrogates. When a SERP contains a diverse set of results, images may be especially useful for determining surrogate relevance at a high-level. Our intuition here is that images may help users quickly understand the diversity of the results as part of the SERP triage process, when trying to identify if their desired query sense is represented at all. Results diversification is a well-established sub-area of IR [1, 5, 6, 21] and is a common strategy in response to an underspecified or ambiguous query. Such queries provide a logical and realistic scenario to study the effect of image-augmented surrogates on diversified result sets. For example, for an ambiguous query such as "jaguar", images might provide an additional advantage in sorting out the results related to Jaguar the car versus jaguar the animal.

We report on two large-scale user studies that investigated the value of image-augmented surrogates. In the first study (Study 1), we focus on relevance judgments at the individual surrogate level. In the second study (Study 2), we focus on relevance judgments at the SERP level. Based on these studies, we address the following research questions:

> **RQ1:** Do images help users make more effective and efficient relevance judgments at the individual surrogate level (Study 1) and at the SERP level (Study 2)?
>
> **RQ2:** To what extent does the "goodness" of the image in reflecting the underlying page content affect the possible benefit (or harm) of including images?
>
> **RQ3:** Do images offer additional benefits at the SERP-level in situations where users need to parse a diversified set of results?

2. RELATED WORK

Augmenting web search result surrogates with a visual component is an intuitively appealing idea, and a variety of approaches have been proposed in the literature. We discuss three main approaches (thumbnails, collages, and single images) and highlight results of several related eye-tracking studies.

Thumbnails—One popular approach has been to use webpage *thumbnails*, a scaled-down bitmap of the webpage as rendered in a browser [3, 23, 25]. Thumbnails are relatively easy to generate and convey layout information about a page that can be helpful both for making relevance decisions [24] and for refinding tasks [13, 23]. Surrogates augmented with thumbnails have been shown to have small benefits over text-only ones in terms of judgement accuracy, but at a slight increase in judgement time [8]. A primary limitation of thumbnails is that they must be fairly large in order to be useful. Kaasten *et al.* [13] found that thumbnails needed to be at least 208 × 208 pixels for website recogntion [13], and Aula *et al.* [3] reported that smaller thumbnails led to poorer relevance judgments. Given their size requirement, many prior studies of thumbnails have not combined them with textual web surrogates (as we do here), but have used SERP layouts customized for thumbnail results [25]. One of our goals in this work was to augment web surrogates with images while keeping the same overall SERP layout.

Collages—A second approach to associating images with web results is to generate "collages" that re-scale and superimpose different features of the page. The goal is to produce a small representation where important features of the page are recognizable and legible. Teevan *et al.* [23] constructed *visual snippets* by combining the title, a salient image, and an important logo from the page. They found benefits of visual snippets for both web search tasks and re-finding tasks. Woodruff *et al.* [25] generated *enhanced thumbnails* which were created by algorithmically modifying the HTML in order to enlarge important features such as headings, pictures, and terms with a high tf-idf weight. Consistent with Teevan *et al.*, Woodruff *et al.* [25] found that different surrogate representations were better for different types of search tasks. Thumbnails and enhanced thumbnails were better for image-based tasks, while text snippets and enhanced thumbnails were better for informational tasks.

Single images—A third approach for adding a visual component to a result surrogate is to include a single image, typically extracted from the underlying web page. This is the type of surrogate enhancement that we consider in this paper. One challenge of this approach lies in automatically selecting an image from the page. Several projects [12, 15, 17] have implemented and evaluated image classifiers designed specifically to identify dominant, content-bearing images from webpages. These efforts have demonstrated that reasonably high accuracy (83%–89%) can be achieved, making automatic image extraction a practical option for selecting images to include in web result surrogates [12, 15]. When a representative images is not present on a page, Jiao *et al.* [12] found that good alternative images could be automatically retrieved from image search engines. Some researchers have used a similar approach of manually gathering images to use in surrogates for the purposes of controlled evaluations [16, 18].

Evaluations of the benefits of using images in surrogates have focused on two main aspects. First are studies that have investigated how well *individual surrogates* help users in making judgments about the content of the underlying webpage. For example, Jiao *et al.* [12] found benefits of using dominant images as surrogates in a content prediction task. Studying image-augmented surrogates in the context of information scent [20], Loumakis *et al.* [16] found that the textual components influenced individual surrogate judgments more than the image did, but noted cases in which 'high-scent' images raised the ratings of a surrogate with 'low-scent' text. Interestingly, when low-scent images were used, participants often ignored the image, suggesting that good images can help, but bad images may not do much harm.

A second group of studies investigated image-augmented surrogates at the SERP-level using information seeking tasks. In this context, Loumakis *et al.* [16] found no differences in effectiveness between text-only surrogates and ones that were image-augmented, but did note strong user preferences

for images. Similarly, Al Maqbali *et al.* [18] found no significant differences in click precision or task completion times of four image-augmentation approaches as compared to a text-only condition. Contrary to these results, Li *et al.* [15] found that tasks done with image-augmented surrogates required considerably fewer clicks and less time to complete than those done with text-only surrogates.

Eye-tracking evaluations—Eye-tracking has been used to understand how users process information on SERPs and how images are used on SERPs [4, 7, 10, 9]. Al Maqbali *et al.* [18] found that users looked at textual components of an image-augmented surrogate more than the image components, but that salient images did draw users' attention. Hughes *et al.* [11] also found that searchers had more eye-gaze fixations on the textual elements of an image-augmented surrogate than on the image, and reported gaze patterns that suggest that participants looked at the text first, and used the image to help confirm or refute the text. Muralidharan *et al.* [19] looked at social annotations integrated into web search results and reported that often participants did not notice small sized images and annotations, but instead focused on URLs and titles in the results. Muralidharan *et al.* also noted that both the location and size of pictures and annotations could have an effect on how much attention they received.

Summary—Taken together, these studies provide a solid context for understanding the role of images in web search surrogates, but leave many open questions. We set out to investigate three unresolved questions in the context of using images extracted from the underlying page to augment a text-based surrogate. First, do the potential benefits from including images manifest at the individual surrogate level, at the whole page level, or both? Second, does the 'goodness' of the image in representing the page have an impact on the possible benefit (or harm) of including images with surrogates? And third, do image-augmented surrogates offer any additional benefits when parsing a diverse set of results on a SERP?

3. METHOD AND MATERIALS

We conducted two user studies to investigate our main research questions. In Study 1, we explore whether images can help users make better surrogate-level relevance judgments. Participants were given a search task and asked to make binary relevance judgments on a sequence of surrogates displayed one at a time. In Study 2, we explore whether images can help users make better SERP-level relevance judgments. Participants were given a search task and a SERP and were asked to find a single web result containing the requested information. The goal for participants was to navigate the SERP naturally by clicking and examining results. Compared to Study 1, Study 2 focused on a more naturalistic setting, where users may not closely examine every surrogate, may be influenced by the other surrogates presented on the SERP, and may be influenced by a surrogate's rank.

3.1 Experimental Variables

Both studies had three experimental variables.

Search task: In both studies, participants made relevance judgments within the context of a search task. Broadly speaking, each *search task* focused on finding information about a particular entity (e.g., "Find information about the Mitsubishi Eclipse."). A set of 150 search tasks were constructed as described in Section 3.2.

Query type: One of our research questions investigates whether image-augmented surrogates offer additional benefits in situations where the results are diverse. We operationalized this by including two queries for each task: an *ambiguous* query that produced a set of diversified results, and an *unambiguous* query that produced a more homogeneous set of results. Our ambiguous queries were intentionally underspecified (e.g., "eclipse") so that they would have a diverse set of search results (e.g., results about "eclipse" the car, the IDE, the astronomical event, and the airplane). The unambiguous query was more specific (e.g., "mitsubishi eclipse") and therefore generated results on a narrower set of query senses. Details about how we generated the queries are explained in Section 3.3.

Image: Our method for augmenting surrogates was to include an image from the underlying web result. We define a 'good' image as one that strongly relates to the main focus of the webpage and a 'bad' image as one that does not. To establish these categories, images were classified manually through a preliminary study, described in Section 3.5. One of our goals was to understand the range of positive and negative influence from augmenting surrogates using good images, bad images, or a mixture of both. Thus, the *image* variable represents the type of image included in the surrogate: *text-only* (title, URL, and summary snippet), *good images* (textual components plus a highly-rated image), *bad images* (textual components plus a poorly-rated image), and *mixed images* (textual components plus a good or bad image chosen randomly).

3.2 Search Task Construction

Search tasks were constructed following a procedure similar to those used in Arguello and Capra [2] and Sanderson [22]. Since part of our goal was to study SERPs with diversified results, we constructed tasks with both an ambiguous and unambiguous query. First, we collected a set of ambiguous entities by identifying Wikipedia disambiguation pages. In Wikipedia, a disambiguation page is used to direct users to articles about different senses of an ambiguous entity. For example, the disambiguation page for "explorer" has links to Wikipedia articles about the Space Shuttle Explorer, the Ford Explorer, and the Internet Explorer browser. A set of 122,130 disambiguation pages were identified using regular expressions. Second, because we wanted to use only entities that might be actually used as a query, we omitted all entities *not* appearing at least once in the AOL query log. This resulted in a subset of 34,151 entities.

Finally, we manually selected 150 of these entities and wrote search task descriptions about one of their senses. In our selection process, we focused on entities with multiple popular senses. Tasks descriptions were written in the form: "Find information about `<entity>`, `<disambiguation>`", for example, "Find information about the Ford Explorer, a model of sport utility vehicle." In this respect, our search tasks focused on finding a webpage about a particular sense of an ambiguous entity. While we did not focus on types of information needs often used in IR studies (e.g., navigational, fact-finding, informational, and transactional), the task of determining whether a search result is about the relevant query-sense is as an important and necessary step in many higher-level search tasks.

3.3 Query Generation

To investigate whether images offer different benefits for SERPs with diversified results, each of our 150 search tasks required both an ambiguous and an unambiguous query. The ambiguous query corresponded to the entity appearing in the Wikipedia disambiguation page title ("explorer"). The unambiguous queries were collected through a preliminary study run on the Amazon Mechanical Turk. Participants were given a search task description and asked to use a live search engine (built using the Bing Web Search API) in order to find a webpage containing the requested information. In the Mechanical Turk, search tasks were presented as Human Information Tasks (HITs). For each HIT, participants were instructed to search naturally by issuing queries and inspecting results. Marking a webpage as containing the requested information concluded the HIT. All HITs were hosted locally on our own server and all user interactions (including all queries issued) were logged. We published a total of 1,500 HITs (150 search tasks $\times$ 10 redundant HITs per task) and priced each HIT at \$0.10 USD.

Given this search interaction data, unambiguous queries were selected in two steps. First, in order to avoid ineffective unambiguous queries, we only considered those queries that were the last query in the session (those which ultimately resulted in the selection of a relevant result). This produced 10 candidate queries per search task. Second, we selected the most common query from each set of 10, breaking ties randomly.

3.4 Text Surrogate Generation

For each combination of search task and query, we collected the top-10 surrogates returned by the Bing Web Search API. Bing returned a title, URL, and query-baised summary snippet (without bolding) for each result. In total, we collected 3,000 textual surrogates (150 search tasks $\times$ 2 query types $\times$ top-10 Bing results). The results were cached in order to use the same textual surrogate components in all experiments.

3.5 Image Selection

Our approach to augmenting web surrogates with images was to include an image pulled from the underlying webpage. We consider a 'good' image as one that strongly relates to the main content of the page and a 'bad' image as one that does not. To study this factor, we felt it was important to use human assessors to establish these good and bad ratings.

To gather the manual judgments, we started by using `wget` to cache a version of each webpage and all the images referenced in the HTML. Most webpages had too many images for manual judgment. In total, we cached about 195,000 images for 3,000 webpages, an average of 65 images per webpage. Therefore, the next step was to identify, for each webpage, a subset of images to be manually assessed. Many images were not suitable for presentation in a surrogate (e.g., decorative images, buttons, unsuitable aspect ratios) and could be filtered. To do this, we created a simple rule-based classifier to select 9 candidate images (maximum) from each page for manual judgement. Our classifier considered features such as aspect ratio, size, color distribution, and image filename. These features are similar to those used in other web image classifiers [15, 12, 17].

The final step was to collect manual judgments on the selected images. Judgments were collected using the Mechanical Turk. The judgment interface displayed the original webpage and its candidate images side-by-side and instructed participants to "rate how well each of the following images 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.

Of the original set of 3,000 pages, 140 were excluded from the final assessment phase because they did not contain any images, or did not contain any images matching our minimum criteria (based on the aspect ratio). We collected five redundant assessments per webpage for a total of 14,300 HITs ((3,000-140) $\times$ 5). For each webpage, the set of images were displayed in random order, with the same ordering used in each redundant HIT. We paid participants \$0.10 USD for each HIT they completed.

Quality control was done in two ways. First, no single worker was allowed to do more than 750 HITs (about 5% of those available). Second, every HIT included a 'check', which corresponded to an image pulled from a different webpage. Assigning this extraneous image a rating of 6 or 7 (on the 7-point scale) was considered a 'failed check'. Participants who failed more than three checks were not allowed to complete further HITs.

After the image ratings were collected, each webpage was associated with a good, bad, and mixed image. We used the mean of the five ratings obtained for each image to represent its overall "goodness". For the good image, we randomly selected an image with a mean rating of five or greater, and for the bad image, one with a mean rating of less than five. Finally, for the mixed image, we randomly selected either the good or bad image. Based on the above criteria, a small number of webpages did not have a good and/or a bad image, and consequently, no mixed image. Missing good, bad, and mixed images were handled differently in Study 1 and Study 2, as described in Sections 3.7 and 3.8.

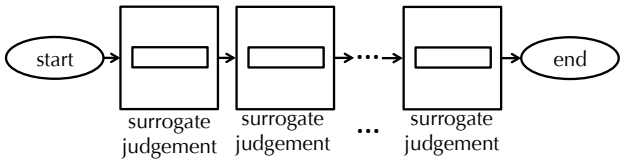
Assessors were largely consistent in their assessment of images. Inter-assessor agreement was computed using weighted Fleiss Kappa (κ_f), which measures chance-corrected agreement between *any* pair of assessors.[1] Because responses were indicated on a 7-point scale, weighted Kappa (using quadratic weights) was used in order to more severely punish disagreements farther on the scale. The Fleiss' Kappa agreement was $\kappa_f = .620$, which is considered *substantial* agreement [14]. While there is some room for improvement, this level of agreement suggests that our assessors were able to perform the image-rating task with acceptable reliability and that the 'good' and 'bad' images used in our experiments were actually good and bad.

3.6 Relevance Judgments

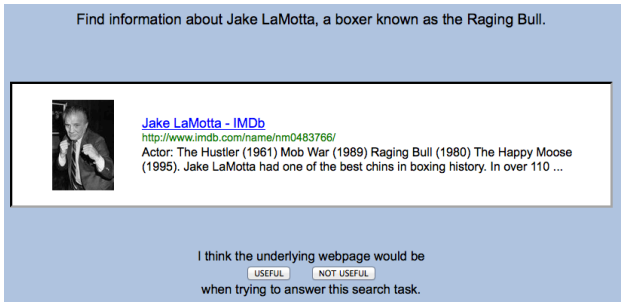
In Study 1 and Study 2, a primary question is whether image-augmented surrogates help users make better judgments about the underlying page. For this analysis, it was necessary to obtain 'ground-truth' relevance judgments at the webpage level.

Two students from our university (not authors) were employed to assess binary relevance. Both assessors judged a common set of 600 of the webpages and one assessor judged the entire set of 3,000. The Cohen's Kappa agreement between assessors was $\kappa = .693$, which is considered *substantial*

[1]Cohen's Kappa was not appropriate because not every assessor judged every image.



(a) flow diagram



(b) surrogate judgment interface

Figure 1: Study 1 flow diagram and example surrogate display.

agreement [14]. Given this high level of agreement, we used the judgements of the assessor who rated all 3,000 pages as our ground-truth.

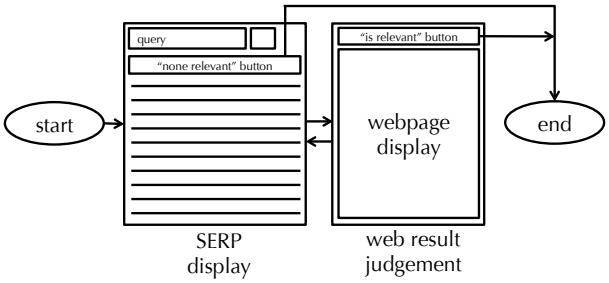
3.7 Study 1

To investigate the effect of images on surrogate-level judgments, Study 1 participants were given a search task and asked to assess a sequence of surrogates. A flow diagram of Study 1 and a screenshot of the surrogate display are shown in Figure 1. Surrogates were presented in sets, one at a time. All surrogates in the same set were associated with the same experimental condition (i.e., the same combination of *search task*, *query type*, and *image*). For each surrogate, the participant was asked whether they thought the underlying webpage would be *useful* or *not useful* for the given search task. Participants could not abstain from making a judgment. The surrogates were formatted to be similar in style to those from commercial search engines. However, all hyperlinks were disabled so that participants had to make their decision using only the surrogate.

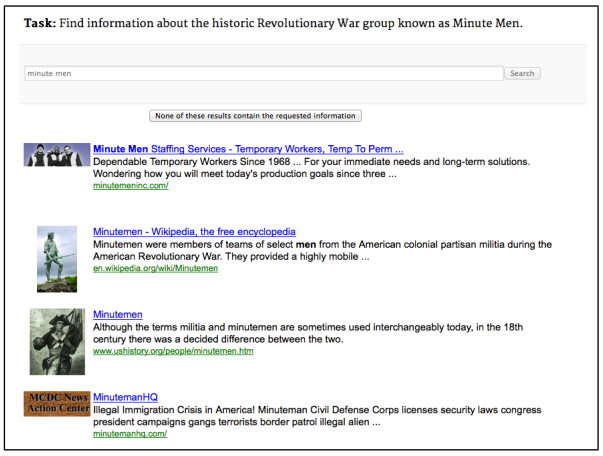
As noted in Section 3.5, not every page had a good, bad, and mixed image available. To ensure an equal number of surrogates in each experimental condition, we removed surrogates without a good, bad, and mixed image, and then removed search tasks with fewer than five surrogates. This resulted in a set of 128 tasks.

Study 1 was run on Mechanical Turk. Each surrogate set was presented to five different participants, resulting in a total of 5,120 HITs (128 search tasks $\times$ 2 query types $\times$ 4 image conditions $\times$ 5 redundant HITs) and a total of 34,000 judgments on surrogates. The order in which the surrogates were displayed was randomized for each HIT to control for possible learning and fatigue effects. Each HIT was priced at \$0.10 USD.

In addition to recording participant judgments for each surrogate, we also recorded the duration of time for which the surrogate was displayed before a response was made. To handle variability in network latency, we used Javascript to pre-load the surrogates on the participant's browser, and



(a) flow diagram



(b) SERP display (cropped)

Figure 2: Study 2 flow diagram and example SERP.

used Javascript and AJAX to record the time duration measures and transmit them to our server in batch mode.

Quality control was carried out by including a 'check' with every HIT. For the check, we presented a surrogate from a different search task. If the participant marked this surrogate as *useful*, we counted this as a 'failed check'. Participants who failed more than three checks were automatically filtered and not allowed to complete further HITs.

Study 1 participants were assigned to experimental conditions randomly, except for two constraints. First, participants were randomly assigned to a single image condition (i.e., image condition was a between-subjects factor). The reason for this was to keep the good, bad, and mixed image conditions consistent and separate. If a participant had been allowed to see both the good and bad conditions, this would be similar to the mixed condition. Second, participants were not allowed to see a search task more than once (even for different query-types).

3.8 Study 2

Study 2 investigated image-augmented surrogates in the context of a whole SERP. In contrast to Study 1, where relevance judgments were made on each individual surrogate, in Study 2 we were interested in simulating a more natural environment, where judgments are made within the context of other surrogates and may be influenced by the user's perception of the overall results and a surrogate's rank.

A flow diagram of Study 2 and a screenshot of the SERP display are shown in Figure 2. Participants were given a search task description and a SERP, which included a query

and its top-10 web results. Participants were instructed to find a web result containing the requested information, or to determine that none did. Participants were only given access to the top-10 results and were not allowed to issue new queries. Clicking on the surrogate title opened the underlying webpage in an HTML frame. Above this frame, a button was displayed that participants could use to indicate that this page was their choice to satisfy the task (Figure 2a, "is relevant" button). Alternatively, participants could use the browser back button to continue searching. Participants were also able to select that none of the results contain the requested information by clicking on a button displayed above the search results (Figure 2a, "none relevant" button). Clicking either button concluded the task.

Like Study 1, Study 2 used three independent variables: *search task*, *query type*, and *image*. As mentioned previously, it was possible for a web result to not have a good, bad, or mixed image. Different from Study 1, in Study 2, if a web result did not have an image for a specific image condition, its text-only surrogate was shown. Study 2 had 1,200 experimental conditions (150 search tasks × 2 query types × 4 image conditions).

Study 2 was also run on Mechanical Turk and had five redundant HITs per experimental condition for a total of 6,000 HITs. Each HIT was priced at $0.10 USD. Quality control was done by comparing a participant's judgments with the ground truth document-level relevance judgments. Participants could make two types of mistakes: they could incorrectly mark a non-relevant result as relevant or could incorrectly click the "none relevant" button when at least one of result was relevant. Either mistake was viewed as a 'failed check' and evidence of careless work. Participants who made more than five mistakes were filtered and not allowed to complete further HITs. Study 2 participants were assigned to experimental conditions randomly, except for the same constraints described for Study 1. That is, participants were assigned to a particular image condition and were not allowed to see the same search task more than once.

4. RESULTS

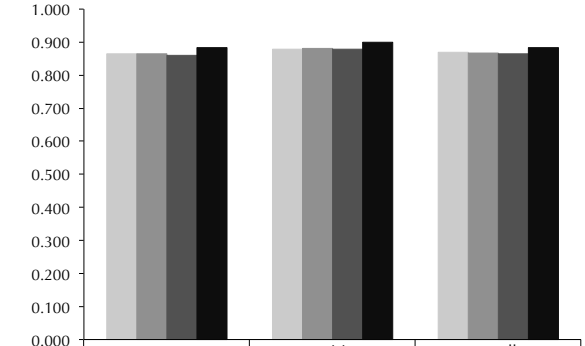
In this section, we present results from our two user studies. In our analyses, when multiple post-hoc tests are conducted, the p-values are adjusted using Bonferroni corrections for non-parametric tests and Tukey-Kramer corrections for ANOVA.

4.1 Study 1 Results

Study 1 was concerned with relevance judgments at the surrogate level. We investigated four outcome measures:

- *Accuracy*, *recall*, and *precision*: Across all the conditions in Study 1, we collected 34,000 relevance judgments on individual surrogates. Comparing these to the document-level judgements, we summed the correct and incorrect judgments and computed aggregated accuracy, precision, and recall measures for each of the four image conditions.

- *Average judgment duration*: For each individual surrogate judgment, we computed the amount of time the participant took to make the judgment. We averaged these individual times for each image condition, again aggregating across tasks and query types.



	accuracy	precision	recall
text-only	0.864	0.879	0.870
bad	0.864	0.882	0.867
mixed	0.861	0.878	0.864
good	0.884	0.900	0.884

Figure 3: Accuracy, recall, and precision measures for each image condition in Study 1.

4.1.1 Accuracy, precision, and recall

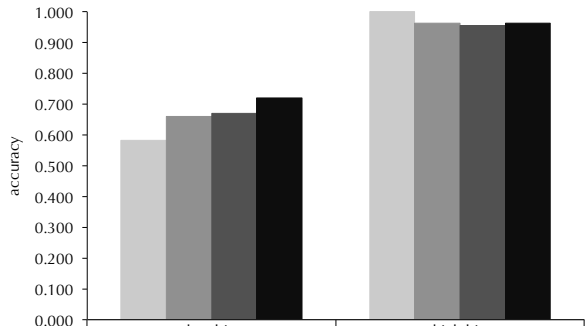
Figure 3 shows accuracy, precision, and recall scores computed across all judgments in each image condition. Using a chi-square test we found a main effect of image condition on accuracy ($\chi^2(3) = 24.26$, $p < .01$). Good images led to 2.3%, 2.7%, and 2.3% improvements in accuracy over text-only, mixed images, and bad images, repectively. While these are only modest improvements, post-hoc comparisons indicated that they were all significant ($p < .01$). No other significant pairwise differences were found.

A similar analysis for precision and recall yielded the same main effects for image condition (precision: $\chi^2(3) = 14.60$, $p < .01$; recall: $\chi^2(3) = 9.97$, $p < .05$). Pairwise comparisons were also consistent, except that for recall, only the improvement of good images over mixed images was significant ($p < .05$).

A few trends from the above results are worth noting. First, accuracy, precision, and recall scores were high, indicating that participants did not have difficulty making relevance judgements at the individual surrogate level. Even the text-only condition had an overall accuracy of .861, suggesting that in many cases the textual elements (i.e., title, URL, and snippet) conveyed the necessary information for participants to make accurate surrogate-level judgements. Second, good images helped participants make significantly better surrogate-level judgements, but these gains were very modest. Interestingly, the small improvements in accuracy from good images came from both precision and recall. This means that good images helped participants identify relevant results *and* reject non-relevant ones. Finally, we did not find significant decreases in accuracy, recall, or precision in the bad or mixed conditions. In other words, while good images helped, bad and mixed images did not hurt. We explore this further in the next section.

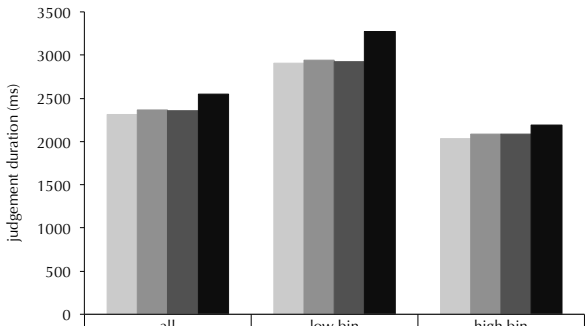
4.1.2 Binned Analysis

Overall accuracy in Study 1 was high, even in the text-only condition. This suggests that most text-only surrogates already provide the necessary information for users to make effective surrogate-level judgements. In this section, we investigate those cases where the text was not as good and the images had a greater chance to help.



	low bin (n=2765)	high bin (n=5735)
text-only	0.583	1.000
bad	0.660	0.963
mixed	0.669	0.954
good	0.721	0.962

Figure 4: Accuracy for the low and high bins in Study 1.



	all	low bin	high bin
text-only	2313	2903	2029
bad	2361	2938	2084
mixed	2357	2928	2083
good	2544	3275	2193

Figure 5: Study 1 average judgment duration.

To facilitate this analysis, we created two bins: a high bin for surrogates that had perfect accuracy in the text-only condition, and a low bin for all others. We placed a surrogate in the high bin if all five of its redundant judgments were correct (relevant/non-relevant) in the text-only condition. Otherwise, we placed it in the low bin. This resulted in 5,735 surrogates in the high bin, and 2,765 in the low bin. The larger high bin is reflective of the overall high accuracy reported above. Our goal in this analysis is two-fold: (1) to investigate whether images helped in situations where the textual surrogate components were less than perfect, and (2) to investigate whether images hurt in situations where the textual surrogate components were already effective.

Figure 4 shows accuracy scores for each image condition for the low bin (left) and the high bin (right). Precision and recall scores are not included in Figure 4 due to space reasons, but are discussed below. First, we consider the low bin. Using a chi-square test of the correct and incorrect judgments, we found a main effect of image condition on accuracy ($\chi^2(3) = 119.67, p < .01$), with post-hoc comparisons indicating that all image conditions were significantly different from each other ($p < .01$), except bad versus mixed. Good images led to a 24% improvement in accuracy over text-only, a 9% improvement over bad images, and an 8% improvement over mixed. Interestingly, even bad images had a 13% improvement over the text-only condition. This improvement may have come from situations where the "bad" image still contained useful information not conveyed by the textual surrogate components. We ran a similar analysis for recall and precision and found the same main effect as accuracy and the same differences in the post-hoc comparisons. These results for the low bin show a benefit of images in situations where the textual components of the surrogate alone did not lead to good relevance judgments.

In the high bin, there was also a significant effect of image condition on accuracy ($\chi^2(3) = 249.42, p < .01$). Compared to the perfect accuracy of the text-only condition (the inclusion criterion for the high bin), bad, mixed, and good images resulted in 4%, 5%, and 4% reductions in accuracy, respectively. Post-hoc tests showed all three differences to be significant ($p < .01$), with no other significant pairwise differences. An analysis of precision and recall for the high bin showed similar results. Results for the high bin suggest that when the textual surrogate components are already good, there is a slight (yet significant) decrease in accuracy from including images irrespective of image quality.

Taken together, our binned analysis shows that images can help low-quality textual surrogates and can hurt high-quality textual surrogates. However, images help the former more than they hurt the later.

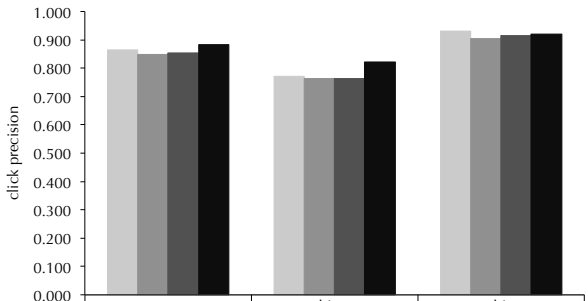
4.1.3 Judgment Duration

We examined the amount of time it took participants to make a relevance judgment. The average judgment durations for each image condition, as computed across all the judgments, are shown on the left side of Figure 5. Out of the 34,000 judgments, we removed 114 outliers with durations greater than 20 seconds. An ANOVA showed a significant effect of image condition on judgment duration ($F(3) = 25.55, p < .01$), with post-hoc tests ($p < .01$) showing that surrogates with good images took longer to judge (2,544ms) than all other conditions (which were each around 200ms less, on average). The middle and right sections of Figure 5 show the average judgment durations for the low and high bins. We found a main effect of bin ($F(3) = 1445.18, p < .01$) with the low bin having a higher average judgment duration than the high bin. Within the low bin, the good images took significantly longer to judge (3,275ms) than all other conditions (each around 300ms less, on average), ($p < .01$). Within the high bin, the good images took significantly longer to judge (2,193ms) than the text-only condition (2,029ms), ($p < .01$). Considered together with the results from the previous section, these findings suggest that good images helped to improve accuracy, but required additional time to process.

4.2 Study 2 Results

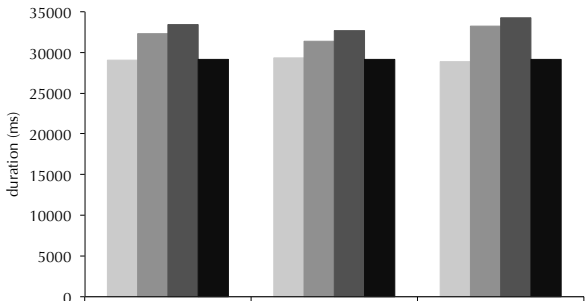
Study 2 examined the effects of image-augmented surrogates at the SERP-level. We focus on two dependent measures for Study 2:

- *click precision*: For each HIT, we computed click precision as the number of clicks on relevant results divided by the total number of clicks made on the SERP.
- *time-to-completion*: For each HIT, we computed the amount of time it took the participant to complete the



	all	ambiguous queries	unambiguous queries
text-only	0.865	0.771	0.931
bad	0.848	0.764	0.905
mixed	0.853	0.763	0.915
good	0.882	0.822	0.920

Figure 6: Study 2 average click precision.



	all	ambiguous queries	unambiguous queries
text-only	29094	29353	28835
bad	32337	31407	33268
mixed	33458	32665	34251
good	29149	29111	29188

Figure 7: Study 2 average time to completion.

task (i.e., to identify and select a relevant result on the SERP or to decide that none of the results were relevant).

Of the 6000 total HITs, there were 5005 where a participant clicked on at least one SERP surrogate. Click precision was computed and analyzed only for these 5005 HITs. In the other 995 cases, the participant clicked only on the "none relevant" button. Overall, the accuracy of pressing the "none relevant" button was high (90.8%) and no difference in accuracy was found for image condition.

4.2.1 Click Precision

Figure 6 shows average click precision for each image condition for all HITs (left), for only the ambiguous queries (middle), and for unambiguous queries (right). A Kruskal-Wallis[2] test showed a significant main effect of query type ($\chi^2(1) = 182.56$, $p < .01$), with the unambiguous queries having a significantly higher average click precision (.918) than the ambiguous queries (.779). The main effect of image condition was marginally significant ($\chi^2(3) = 7.54$, $p = .06$), with post-hoc pairwise tests showing that good images were slightly better (.882) than bad images (.848), ($\chi^2(1) = 2.55$, $p = .06$). Good images also had higher click precision than the text-only condition (.865), but this difference was not found to be significant.

For ambiguous queries, the effect of images was marginally significant ($\chi^2(3) = 6.80$, $p = .08$). As can be seen in the middle section of Figure 6, good images had higher average click precision (.822) versus text-only (.771), bad (.764), and mixed (.763), but the post-hoc pairwise tests did not find significant differences (good versus mixed had a Bonferroni-adjusted p-value of .11). For unambiguous queries, no effect of images was found ($\chi^2(3) = 4.26$, $p = .23$).

These results show that overall, good images were not significantly different from the text-only condition in terms of click precision, but were slightly higher than the bad or mixed images. For the ambiguous queries, good images had a noticeable improvement over all other conditions (including text-only), but these differences were not enough to show statistical significance.

[2]Since the data for click precision was non-normally distributed, we used non-parametric tests.

4.2.2 Time-to-completion

Figure 7 shows the average time-to-completion for each image condition for all HITs (left), the ambiguous queries (middle), and the unambiguous queries (right). Across all HITs, image condition had an effect on time-to-completion ($F(3) = 9.83, p < .01$), with post-hoc tests showing that participants took significantly more time with bad and mixed images than with text-only and good images ($p < .01$). Text-only and good images were statistically indistinguishable ($p = 1.0$).

The same general trend was observed for both query types. For the ambiguous queries, image condition had a significant effect on time-to-completion ($F(3) = 3.03, p < .05$), with post-hoc tests showing that mixed images required (marginally significant) longer times than text-only ($p = .08$) and good images ($p = .05$). For the unambiguous queries, image condition also had a significant effect ($F(3) = 7.23, p < .01$), with post-hoc tests showing that bad and mixed images required longer times than text-only and good images ($p < .05$), and that text-only and good images were statistically indistinguishable ($p = .99$).

These results indicate that there was an increase in time-to-completion from bad and mixed images, but not from good images. With good images, participants were able to complete the task of finding a relevant result on the SERP (or determine that none of them were relevant) as fast as with text-only surrogates.

5. DISCUSSION

In this section, we review our results in the context of our three research questions and in relation to prior work.

RQ1 - Do image-augmented surrogates help relevance decisions? Results from Study 1 show that when making individual surrogate judgments, good images provided only a 2.4% increase in accuracy as compared to the text-only condition, and led to slight increases in judgment duration. In Study 2, for the SERP-level search tasks, good images had a 2.0% increase in click-precision over text-only, but this difference was not significant. Interestingly, in Study 2, good images did not increase search task completion times as compared to text-only, but bad and mixed images did cause users to take more time. Across both studies, we find

that the improvements from adding images are small, and are contingent on the ability to identify a good image to represent the page.

We interpret our results to be consistent with several recent studies. Both Al Maqbali *et al.* [18] and Loumakis *et al.* [16] found no significant benefits of image-augmented surrogates over text-only ones in terms of effectiveness or efficiency. Dziadosz and Chandrasekar [8] reported benefits of images, but their gains were also quite small (2%-3%). These studies and ours used different methods and different tasks, yet the results largely agree—adding images to web results surrogates has very little effect on effectiveness and efficiency *in the general case*. As we will dicuss below, larger benefits are possible for special sitautions. In contrast to these results, in two small studies, Li *et al.* [15] found substantial gains from image-augmented surrogates compared to text-only ones for information seeking tasks. The reason for these conflicting results is not clear, but may be due to differences in the tasks or experimental protocols used.

One area where we found particular benefits of image-augmentation was in cases where the textual components of the surrogate did not provide sufficient information. In our 'binned' analysis of Study 1, we found that for surrogates where the text alone did not have perfect judgment accuracy (the 'low-bin'), adding good images increased accuracy by 24%. This result is consistent with results from Loumakis *et al.* [16] that 'high-scent' images can help improve surrogates with 'low-scent' text, and observations from Hughes *et al.* [11] that images are sometimes used to help confirm or refute the textual parts of a surrogate. In a study of social annotations embedded into SERP results, Muralidharan *et al.* [19] noted that when summary snippets were shorter, the social annotations were noticed more. This may also help to explain the the 'low-bin' results—users looked to the images as another source of information. Together, these results suggest a possibly beneficial use of image-augmentation in cases where the text of a surrogate is lacking. The idea would be to predict cases where the text might not be sufficient and to then augment the surrogate with an image. There are a number of features that might be useful to predict text surrogates that would benefit from images (e.g., short or disjointed summary snippets, uninformative titles). We see this as an interesting area of future work.

RQ2 - How does the 'goodness' of the image affect its benefits? Results of Study 1 show that good images had higher accuracy than text-only, bad, and mixed conditions, and that bad and mixed had no difference compared to text-only. Good images took around 200ms longer to judge, but the other conditions were statistically similar to each other. The binned analysis showed that for the low bin, all images helped, but that good images helped more (and resulted in longer judgment times). For Study 2, the results were somewhat different. Good images had small (non-significant) increases to average click-precision versus text-only, but with no increase in total task completion time. Bad and mixed images, however, had no click-precision differences versus text-only, but did increase the task completion time. Our interpreation of these results is that there are differences between using good, bad, and mixed images in surrogates in terms of effectiveness and efficiency measures.

When evaluating an individual surrogate, bad and mixed images appear to have been largely ignored. In Study 1, they were no different than text-only in terms of judgment accuracy or time. Loumakis *et al.* [16] found similar results for 'low-scent' images in their study, noting that participants mostly ignored these non-informative images. In Study 2, bad and mixed images had a negative effect on task completion time, suggesting that they confused or slowed down users' triage of the SERP. Both Loumakis *et al.* [16] and Dziadosz and Chandrasekar [8] commented on the possibility of bad images to confuse or mislead users, and we see this as an area of potential risk in choosing to augment surrogates with images from the page.

We also note that our 'good' images were selected based on multiple human assessors' ratings of how well the images reflected the main content of the page. We used this approach to establish an upper-bound on the level of improvement that could be expected from using an image from the underlying page. Thus, in terms of effectiveness, even at their best, images only help a little.

Across our two studies, the observed improvements from images are small, and are contingent on the ability to identify a good image to represent the page. Work by Li *et al.* [15] has shown that it is possible to construct efficient and robust classifiers to identify salient images on a page, but the accuracy of these clasffieres is in the 85% range, so some bad and mixed images will be selected. In addition, not all pages contain good images [12]. System designers would need to determine if the modest benefits we report are worth the increased overhead and tradeoffs. Of course, there may be additional reasons beyond effectiveness and efficiency that would influence designers to include images. Loumakis *et al.* [16] reported a strong subjective user preference for combined image-augmented surrogates, and Jiao *et al.* [12] found that images extracted from the underlying page were a preferred summarization method.

RQ3 - Do images help more for SERPs with diversified results? Our third research question (RQ3) considered whether the effects of images might be different in situations where the results were diverse. We operationalized this by examining SERPs generated from both ambiguous queries (diversified results) and unambiguous queries (more homogeneous results). Study 2 addressed this question and found an interesting trend. For the ambiguous queries, good images had a 7.0% higher click precision compared to the text-only condition. While this difference was not statistically significant, it was a noticeable increase in Figure 6 and was not present in the umambiguous queries. This suggests that images may help ambiguous queries more than the unambiguous ones. We see this as another situaion in which image-augmented surrogates could be selectively applied to increase users' ability to triage a SERP and make good relevance decisions. In cases where an ambiguous query was issued, or in which the search engine has decided to present diversified results, surrogates could be augmented with images to help users interpret and make sense of the results presented on the SERP.

Implications—Although in the general case, we found only small benefits of augmenting web search results with images from the underlying page, we identified several situations where images helped measurably. Image-augmentation could be *selectively* applied when the textual components are poor and when the search results are diversified.

6. CONCLUSIONS

We presented two user studies that investigated whether image-augmented web surrogates help users make better relevance judgements at the individual surrogate level (Study 1) and at the SERP level (Study 2). Our results showed *very small* improvements from augmenting surrogates with "good" images. At the individual surrogate level, good images improved judgement accuracy by 2.3% over text-only surrogates. At the SERP level, good images improved click precision by 2.0% over text-only surrogates.

While good images provided only small improvements overall, both studies found larger improvements in special cases. At the individual surrogate level, good images provided a 24% improvement in accuracy for those text-only surrogates that did not elicit quality judgements from participants. At the SERP level, good images provided a 7.0% improvement in click precision when the query was ambiguous and the results were diversified.

Taken together, the above results suggest that augmenting web result surrogates with images indiscriminately seems risky. Our "good" images were *manually* selected using redundant assessors and were meant to represent the best possible outcome from an algorithmic extractor. We outline two promising directions to augment web result surrogates *selectively*, based on the quality of the textual surrogate components (as suggested by Study 1 results) and based on the diversity of the results presented on the SERP (as suggested by Study 2 results).

7. ACKNOWLEDGEMENTS

We thank Emily Vardell and Wan-Ching Wu for creating document relevance judgements.

8. 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 and R. Capra. The effect of aggregated search coherence on search behavior. In *CIKM 2012*, pages 1293–1302. ACM, 2012.

[3] A. Aula, R. M. Khan, Z. Guan, P. Fontes, and P. Hong. A comparison of visual and textual page previews in judging the helpfulness of web pages. In *WWW 2010*, pages 51–60. ACM, 2010.

[4] G. Buscher, E. Cutrell, and M. R. Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *CHI 2009*, pages 21–30. ACM, 2009.

[5] 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.

[6] 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.

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

[8] S. Dziadosz and R. Chandrasekar. Do thumbnail previews help users make better relevance decisions about web search results? In *SIGIR 2002*, pages 365–366. ACM, 2002.

[9] L. A. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR 2004*, pages 478–479. ACM, 2004.

[10] Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *CHI 2007*, pages 417–420. ACM, 2007.

[11] A. Hughes, T. Wilkens, B. Wildemuth, and G. Marchionini. Text or pictures? an eyetracking study of how people view digital video surrogates. In *Lecture Notes in Computer Science, Volume 2728*, pages 271–280. Springer, 2003.

[12] B. Jiao, L. Yang, J. Xu, and F. Wu. Visual summarization of web pages. In *SIGIR 2010*, pages 499–506. ACM, 2010.

[13] S. Kaasten, S. Greenberg, and C. Edwards. How people recognize previously seen web pages from titles, urls and thumbnails. In *HCI 2001*, pages 247–265. ACM, 2001.

[14] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[15] Z. Li, S. Shi, and L. Zhang. Improving relevance judgment of web search results with image excerpts. In *WWW 2008*, pages 21–30. ACM, 2008.

[16] 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.

[17] T. Maekawa, T. Hara, and S. Nishio. Image classification for mobile web browsing. In *WWW '06*, pages 43–52. ACM, 2006.

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

[19] A. Muralidharan, Z. Gyongyi, and E. Chi. Social annotations in web search. In *CHI 2012*, pages 1085–1094. ACM, 2012.

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

[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] J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu. Visual snippets: summarizing web pages for search and revisitation. In *CHI 2009*, pages 2023–2032. ACM, 2009.

[24] A. Tombros, I. Ruthven, and J. M. Jose. How users assess web pages for information seeking. *JASIST*, 56(4):327–344, 2005.

[25] A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrsion, and P. Pirolli. Using thumbnails to search the web. In *CHI 2001*, pages 198–205. ACM, 2001.