A Comparison of the Use of Text Summaries, Plain Thumbnails, and Enhanced Thumbnails for Web Search Tasks

Allison Woodruff,[1] Ruth Rosenholtz, Julie B. Morrison, and Peter Pirolli
Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304.
E-mail: woodruff@parc.xerox.com.

Andrew Faulring
Xerox Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304 and Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

We introduce a technique for creating novel, enhanced thumbnails of Web pages. These thumbnails combine the advantages of plain thumbnails and text summaries to provide consistent performance on a variety of tasks. We conducted a study in which participants used three different types of summaries (enhanced thumbnails, plain thumbnails, and text summaries) to search Web pages to find several different types of information. Participants took an average of 67, 86, and 95 seconds to find the answer with enhanced thumbnails, plain thumbnails, and text summaries, respectively. As expected, there was a strong effect of question category. For some questions, text summaries outperformed plain thumbnails, while for other questions, plain thumbnails outperformed text summaries. Enhanced thumbnails (which combine the features of text summaries and plain thumbnails) had more consistent performance than either text summaries or plain thumbnails, having for all categories the best performance or performance that was statistically indistinguishable from the best.

Introduction

Internet users spend a significant amount of time examining search engine results; one commercial search engine vendor claims to answer 40 million search queries each day (Technology Review, 2000). The user must page through lists of Web documents, briefly evaluating each for possible relevance to a particular information need. Improving the efficiency of this tedious process directly benefits the end-user and, by improving end-user satisfaction, indirectly benefits the search engine vendor.

The search engine can increase user efficiency by (1) returning higher-quality document lists (e.g., through better index coverage and ranking algorithms), or by (2) providing information that allows the user to evaluate the results more quickly and accurately. Search engine vendors attack both problems. The standard practice with regard to approach (2) is to provide brief textual summaries of the Web documents. We believe that this latter practice can be improved upon.

We have performed a quantitative comparative study of textual and graphical summarization mechanisms applied to search engine results. We argue that graphical summaries of the documents—thumbnail images—can greatly increase the efficiency by which end-users process search engine result sets. For example, thumbnails allow users to classify a Web page's genre very rapidly. Most interestingly, our empirical results suggest that, if properly designed, *enhanced thumbnails* (thumbnails augmented with readable textual elements) deliver the efficiency benefits of both text summaries and *plain* thumbnails (graphically scaled versions of documents).

To understand why this might be the case, one must understand the relative advantages and disadvantages of presenting information in textual and graphical form. We now turn to a brief discussion of the relative tradeoffs, with particular attention paid to the specific application of Web search results.

Text summaries are terse but are verbal rather than visual. They require little storage space and can, therefore, be downloaded quickly. Additionally, text summaries often contain a great deal of valuable information about each document. For example, search engines commonly provide the document's URL, title, size, and a few phrases or sentences that either summarize the document or emphasize some of the search keywords. On the other hand, text summaries do not provide much information about the page layout or any image contained in the page. Furthermore, the

Published online XX Month 2001 • DOI: 10.1002/asi.0000

user must read the text summary. Reading lists of search results is tiring, and empirical studies show that the average search engine user is unwilling to read through more than a few pages of such listings (e.g., Jansen, Spink, Bateman, & Saracevic, 1998).

Simple graphical summaries have strengths and weaknesses that are complementary to those of text summaries. As images, thumbnails typically require more storage space than text summaries, and therefore, they generally download more slowly than text summaries. Textual content in plain thumbnails is less accessible than that in text summaries, as it is difficult to read and is not conveniently summarized. However, graphical summaries do provide information about the layout, genre, and style of the page. If the user has previously seen the page, or one like it, the visual representation may aid in recognizing or classifying it. This becomes even more compelling in view of the fact that the human visual system can process images more quickly than text. Graphical information can speed many tasks tremendously. We can get the "gist" of an image in 110 ms or less, changing fixation roughly every 300 ms (Coltheart, 1999). In that same 110 ms, we can read on average less than 1 word or skim two words. (The average reader of English reads about 4.2 words per second and can skim or scan at roughly 17 words per second) (Chapman, 1993). Furthermore, searching for a *picture* of a particular object among other pictures is faster than searching for the *name* of that object among other words (Paivio, 1974).

In this article we compare text summaries with plain thumbnails (simple reduced-size images), as well as with enhanced thumbnails, which we have designed in the hopes of capturing some of the advantages of both text summaries and plain thumbnails. We make several contributions:

(1) Our enhanced thumbnails consist of a reduced image of the document along with various forms of emphasis of information in the document. Previous work has generally presented plain thumbnails (e.g., Ayers & Stasko, 1995; Hightower, Ring, Helfman, Bederson, & Hollan, 1998; Kopetzky & Mühlhäuser, 1999; Robertson, Czerwinski, Larson, Robbins, Thiel, & van Dantzich, 1998), and/or represented properties of the document in an abstract form (e.g., Cockburn, Greenberg, McKenzie, Jasonsmith, & Kaasten, 1999; Wynblatt & Benson, 1998). Our enhanced thumbnails enforce readability of certain parts of the document within the thumbnail and display highlighted keywords transparently overlaid on the reduced document.
(2) Much of the previous work on thumbnails has emphasized using them for recall of previously seen documents. By contrast, we focus on using them in an application in which the user is unlikely to have seen many of the documents before.
(3) We present a study comparing the effects of text summaries, plain thumbnails, and enhanced thumbnails on realistic search tasks. Users had better and more consistent performance when using enhanced thumbnails than when using the other summary types.

While some of the above issues are addressed in our previous paper as well (Woodruff, Faulring, Rosenholtz, Morrison, & Pirolli, 2001), this article contains a more in-depth discussion of our design, and introduces new results and further interpretation of our findings.

In the next section we discuss related work. In the subsequent sections, we discuss our system for generating thumbnails, our study to compare text summaries with plain and enhanced thumbnails in a search task, and future work and conclusions.

Related Work

Previous work includes several different designs for thumbnails. A number of programs generate plain thumbnails. These include many graphical editors, recent versions of Microsoft® Windows®, and the systems described by (Hightower, et al., 1998; Kopetzky & Mühlhäuser, 1999; Robertson et al., 1998), among others. Ayers and Stasko's (1995) thumbnails are similar to plain thumbnails, consisting of a reduced view of the upper left corner of a document.

Other programs generate more complex thumbnails. Cockburn, Greenberg, McKenzie, Jasonsmith, and Kaasten (1999) generate thumbnails that consist of reduced images plus "dogears" that indicate bookmarked and frequently visited pages. Helfman (1999) selects representative images from a document and creates reduced scale images of these to serve as a thumbnail for that document. Wynblatt and Benson (1998) produce Web page "caricatures." These caricatures contain select features of a page, often rendered in an abstract form: title, representative image, number of images, abstract, etc. These caricatures do not preserve layout and lack some of the visual information that might be naturally available in a reduced scale image of the page. For example, rather than having the user judge link density of a Web page from an image of the page, this density is represented by the background color of the caricature.

TileBars (Hearst, 1995) are abstract representations of documents that graphically indicate the text segments in which search terms appear. Our enhanced thumbnails show the relationship among occurrences of search terms in the context of the document, and at a finer granularity than TileBars do. However, the enhanced thumbnails do not provide as compact an overview of the relationship between search terms as TileBars.

A number of systems employ thumbnails. Although a small number of companies have recently introduced software for using plain thumbnails to search the Web,[1] most of the previous work in this area involves previously viewed documents, in the hope that a thumbnail preview may help the user's memory and thus aid in the task. A commonly considered task is navigation through previously viewed Web pages (e.g., Ayers & Stasko, 1995; Card, Robertson & York, 1996; Cockburn, et al., 1999; Hightower et al., 1998; Fn1

[1] e.g., http://www.room102.com.

Robertson et al., 1998). In addition, a number of systems use thumbnails to aid in the management and retrieval of files on a user's computer, tasks for which it is reasonably likely that the user already would have seen the document or image represented by the thumbnail. Graphical editors, for instance, allow the user to preview an image or a collection of images. Recent versions of Microsoft® Windows® provide a thumbnail view of documents (e.g., HTML or image formats) within a folder.

Kopetzky and Mühlhäuser (1999) describe a system that shows previews of target Web pages: when the user moves the mouse cursor over a link in a Web page, a thumbnail of the target page appears temporarily. Though in many cases the user would not have previously seen the documents represented by these thumbnails, the authors justify the use of thumbnails as a memory aid.

In addition to creating applications that use thumbnails, researchers have studied the utility of thumbnails in a memory task. Czerwinski, van Dantzich, Robertson, and Hoffman (1999) asked users to spatially lay out 100 Web pages in Data Mountain, and then measured their performance at retrieving those documents a few months later. After a brief learning period, users were just as good at retrieval whether thumbnails were present, or only plain white boxes representing the documents. This might suggest a lack of utility for thumbnails, but the study may underestimate the importance of thumbnails, as users saw their layout with thumbnails present repeatedly throughout the study. Interestingly, users subjectively ranked the thumbnail images as the most helpful feature for retrieval.

This focus on thumbnails as an aid to memory in retrieving previously seen documents leads us to ask whether thumbnails are useful only when the user has already seen the corresponding documents. In this article we examine the use of thumbnails in a Web search task, in which few, if any, of the documents are likely to have been previously viewed.

System

We implemented a system that generates both plain and enhanced thumbnails of HTML documents. The tool is written entirely in Java, and utilizes a component Web browser, ICE Browser (Wind River, 2000). The component browser provides access to the document as both an HTML document (source form) and a graphics object (rendered form). As we will see, having convenient access to both interfaces greatly simplifies the internal structure of the system.

Our system works in three stages. First, the preprocessor modifies the HTML in the original page, for example, to change the color or size of certain elements. Second, the renderer creates a scaled version of the modified HTML. Third, the postprocessor modifies the image output by the renderer, for example, to reduce its contrast or to add text callouts. This architectural separation is due to the fact that the various transformations are most easily applied to the document in different intermediate formats. The system requires on the order of a few seconds to generate a thumbnail from the raw HTML, not counting network latencies. In this section, we describe these stages in turn. We then discuss some design issues that cut across the stages.

HTML Modification

After retrieving the HTML document associated with a given URL, the preprocessor adjusts the appearance of the HTML elements. The user specifies the desired adjustments using an associative list of phrase/style pairs (or tag/style pairs). For example, the user might specify that each instance of the word "recipe" should be highlighted in "yellow." Similarly, the user can specify that the text of each H1 header tag should be a certain size. Compare the plain thumbnail in Figure 1a with the modified thumbnail in Figure 1b. (Note that the examples of thumbnails presented in the article and used in the experiment show only the top of the Web page, if it is a long document. The system also allows the generation of a thumbnail of the full document.) F1

This functionality is supported as follows. ICE Browser implements portions of the W3C Document Object Model (DOM) Level 1 Specification (World Wide Web Consortium, 1998b), a standard interface for programmatically accessing and modifying HTML documents. The DOM presents the document as a hierarchy of HTML elements, with each element having an associated Cascading Style Sheet (CSS) style definition (World Wide Web Consortium, 1998a). We can modify the HTML document's appearance by manipulating each element's CSS style.

One particularly useful modification is to adjust an element's font size such that the text is still "readable" in the thumbnail, where readability is specified as a given font size in the final rendered image. Compare the header text in Figure 1b with the header text in Figure 1a for an example of making an element readable.

Rendering

This component delegates the rendering of the (modified) HTML to ICE Browser. Because ICE Browser uses the Java2D interface, the scaling factor for the entire document can be specified by a single operation on a graphics context object.

Image Modification

The postprocessor implements a variety of transformations that cannot be expressed in HTML. For the most part, these transformations require some amount of image processing. For example, a color wash may be applied, or additional graphical elements may be overlaid onto the thumbnail.

One useful modification is to render text phrases as *callouts* (enlarged text overlays) on top of the original thumbnail. The system accepts a phrase, a scale factor at

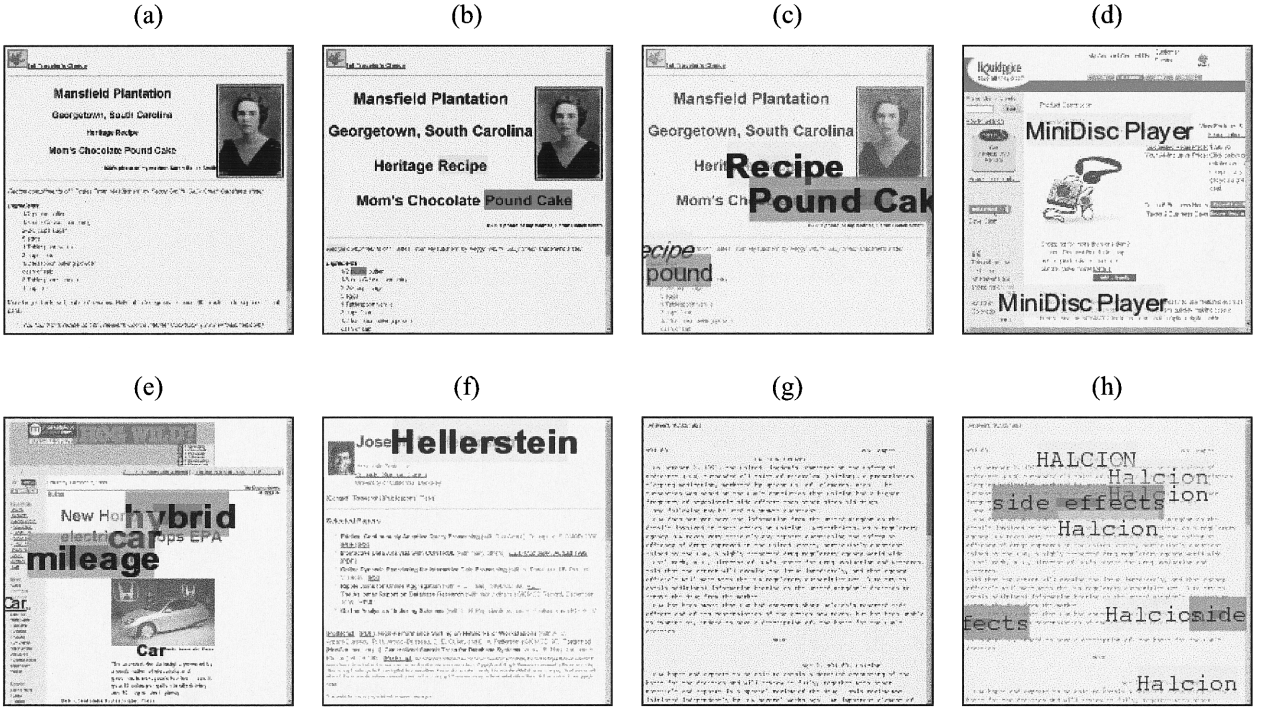

FIG. 1. (a) Plain thumbnail. (b) Thumbnail enhanced with HTML modification. (c) Thumbnail enhanced with HTML and image modification. (d) E-commerce genre example. (e) News genre example. (f) Homepage genre example. (g) Plain thumbnail of textual page. (h) Enhanced thumbnail of textual page.

which to rerender the phrase, and an alignment parameter for positioning the callout relative to the original position of the phrase within the document. The resulting transformation can be easily applied to a specified subset of elements using Java2D interfaces. For example, in Figure 1c, the phrase "Pound Cake" was rendered center-aligned over its original position at four times its original size.

Sometimes the phrases spill outside the edge of the thumbnail, such as the phrase "Pound Cake" in Figure 1c. The system supports various spills rules, allowing the callout to extend over the edge of the reduced page image, so that no clipping occurs, or specifying that the original thumbnail size be retained, clipping any spillage.

Design Issues

The discussion above provides an architectural view of the system, and does not capture the many individual decisions involved in its design. These decisions often required significant attention to visual perception and attention management issues. As an example, in this subsection, we focus on three different techniques we use to manage the user's attention and interpretation of the enhanced thumbnails.

HTML modification of elements. We experimented with a number of ways of modifying HTML to try to draw attention to certain keywords and phrases, for example, dramatically changing the font size, text color, or background color of certain textual elements in the page. However, observations of a large number of thumbnails indicated that because HTML documents have such diverse fonts, colors, and designs, such changes most often appear as though they occur in the original document. A colored text header or an enlarged word generally look like they were created by the original HTML author, not like elements that we have chosen to emphasize after the document was authored. Because these modified textual elements are effectively "grouped" with the original document, they do not draw attention as effectively as they might if they appeared to lie in a separate visual layer, "on top" of the original document.

We conclude that HTML modification is not appropriate for emphasizing elements such as keywords. However, we observe that these modifications are highly appropriate for modifying text that we would like to make readable without explicitly drawing attention to it, for example, text headings. We find that enlarging the size of the headings in the HTML greatly increases the utility of the thumbnail, but the change is so subtle that users often take advantage of the feature without being consciously aware that the text has been enhanced.

Visual layering. One effective way to draw attention to elements is to put them in a separate visual layer. When elements are in a separate visual layer from the original document, they seem to "pop out," thereby drawing the user's attention. Evidence suggests that a user can selectively attend to different layers defined by transparency

(e.g., Lankheet & Verstraten, 1995) and other depth cues (e.g., Hoffman & Mueller, 1994).

We experimented with a number of ways of modifying the image after it was rendered to create callouts that appear to be in a separate visual layer. We first highlighted text overlays with an opaque highlight color, but found that opaque overlays tend to occlude much of the thumbnail, making it difficult for the user to extract gist. Alpha-blending the overlay highlight color with the original thumbnail to create a transparent overlay occludes less of the page, and in addition provides a strong cue that the overlays are additions to the pages, as opposed to being mark-up included by the original author. In our experiments, we found an alpha value of 0.5 to give good results.

Visual layering is appropriate for elements to which we wish to draw attention, for example, keywords.

Color management. We were particularly interested in creating readable, attention-grabbing callouts of keywords. For the dark text common to many Web pages, light, unsaturated background colors most facilitate reading. However, generally speaking, the more saturated a color is relative to surrounding colors, the more it tends to draw attention. We deal with these conflicting requirements for the overlay highlight colors in two ways. First, we wash the entire original thumbnail with a white, transparent fill (we used an alpha of 0.4). Notice the difference between Figure 1a and Figure 1c. This effectively desaturates the original thumbnail. Because to draw attention a color needs to be saturated *relative to surrounding colors,* desaturating the thumbnail allows us to get the same attention-grabbing results with less saturated highlight colors.

After desaturating the original thumbnail, we then used a model of visual search (Rosenholtz, 1999) to select highlight colors that were just saturated enough to "pop out" against a typical thumbnail from our corpus. The resulting highlight colors greatly resemble those colors actually found in highlight pens.

An added benefit of using transparent highlight colors is that this process also works for light text, so long as that text was easily readable in the original document. To be readable, the light text must have occurred against a dark background in the original document. When overlaid against a dark background, our transparent highlight colors produce a dark highlight color against which light text is easily readable. The darkness of the background does not affect the saturation of the highlight color, and thus the highlight will still tend to draw attention.

By combining these image modification techniques, we are able to create callouts that can be easily detected while skimming, while simultaneously allowing the user to get the gist of the underlying thumbnails. In fact, it may be possible for the user to get the benefits of the callouts without needing to actually read much of the highlighted text. In our specific design, we highlight a given word with the same color in each thumbnail, so the user may make use of the fact that all overlays of that color correspond to a particular word.

Design of an Experiment to Compare Different Summary Types in a Search Task

We designed our enhanced thumbnails in the hope of capturing advantages of both text summaries and plain thumbnails. In this section, we discuss a study that we conducted to compare text summaries, plain thumbnails, and enhanced thumbnails. Participants in this study performed tasks that were much like typical Web search tasks. In addition to examining the difference in performance among the three summary types, this task allows us to test whether thumbnails are useful for a task in which the user has never seen the documents represented by the thumbnails. (Recall that, in the past, thumbnails were typically used to aid in recall of documents already seen by a user.)

Participants

Data were collected from 18 members of the Xerox PARC community, 6 women and 12 men. Participants ranged in age from 19 to 56, with a mean age of approximately 35. All were experienced Web users, and reported using the Web daily. All were familiar with the Microsoft® Internet Explorer browser, and most reported using it as their primary browser. Each participant reported using search engines to find information on the Web.

We acknowledge that our participants are not representative of the typical Web user. However, this simply makes our experiment a conservative test of how well enhanced thumbnails would perform against text summaries. First, these are individuals who have a great deal of experience searching for information using the text summaries provided by search engines. Given that they are novice users of thumbnail searching, better performance in the thumbnail conditions would imply that the thumbnail design provides a greater search advantage than both search experience and text summary design. Second, if participants perform better at inexperienced thumbnail search than at experienced text search, this suggests that novice Web users will be able to use thumbnails more effectively than text summaries. This idea was underscored by one participant who volunteered that using enhanced thumbnails was intuitive and should facilitate novice search. To verify these expectations regarding expert versus novice search, future studies will investigate the use of thumbnails by less experienced Web searchers.

Question Categories

We chose four different question categories and developed three questions within each category. First, participants were asked to locate a picture of a given entity. Second, participants were asked to locate the homepage of an individual whom they did not know. Third, participants were asked to locate a consumer electronics item for pur-

TABLE 1. Categories of questions performed by participants.

Category	Characteristics	Example question	Approx. # answers
Picture	Requires identification of a graphical element	"Find a picture of a giraffe in the wild."	8/100
Homepage	Requires genre classification (correct pages somewhat textual, many incorrect pages entirely textual)	"Find Kern Holoman's homepage."	1/100
E-commerce	Requires genre classification (correct pages highly graphical; incorrect pages highly graphical, e.g., product reviews)	"Find an e-commerce site where you can buy a DVD player. Identify the price in dollars."	15/100
Side-effects	Requires semantic information (word proximity and position in layout useful, genre useful)	"Find at least three side effects of Halcion."	20/100

chase. Fourth, participants were asked to locate three or more side-effects of a given drug. Table 1 contains example questions from each category. In addition to the questions in these four categories, we developed another six practice questions, for example, "Find the mileage of a hybrid car."

The nature of the categories affected the number of possible correct answers within the set of 100 pages. People typically have a single homepage, as did the individuals asked about in our study. In contrast, numerous sites contain pictures, products, and medical information. The question categories used in this study differed in the percentage of pages containing an acceptable answer, but within each question category there were approximately the same number of answers (see Table 1). Despite this difference, the number of acceptable answers did not correlate with participants' solution times. The Side-effects category, which contained the most possible answers, had the slowest time, while the Homepage category, with the least answers, had only the second fastest time. As will be made clear in the Results section of this article, response times were related to the content of the pages, and not to the percentage of acceptable answers.

The questions in these categories represent tasks users commonly perform on the Web. Morrison, Pirolli, and Card (2001) have developed a taxonomy of user tasks based on an analysis of over 300 Web users' comments about what Web activities significantly impacted their decisions and actions. The three most common task types were e-commerce (21%), medical (13%), and finding people (9%). Morrison and colleagues' data includes only information that led to a significant action or decision. We included the picture category because we believe it is representative of a common but less "significant" class of queries: searching for graphical content such as photographs or maps. The query results for our question categories yield Web pages that are both semantically and visually different. See the "Characteristics" column in Table 1.

Materials

We constructed our materials for the study in three phases: (1) we archived the Web pages; (2) we created text and thumbnail summaries of the archived Web pages; and (3) we created HTML pages that showed collections of summaries.

Archiving web pages. Our corpus is based on URLs extracted from search results from Google.[2] As an example, for the e-commerce question on DVD players, we programmatically queried Google using the terms "DVD" and "player" and extracted URLs from the result pages. Because the contents of Web pages often change, we downloaded the pages associated with these URLs to create a consistent set of Web pages to show to our participants. Storing the pages locally provides the added advantage that network delays are avoided, allowing for more consistent response times.

Creating summaries. After downloading the pages, we created three different summary materials for each page. First, we extracted the Google *text* summary associated with each URL. These summaries include the page's title, excerpted text with search terms shown in bold, and the URL. Second, we created a *plain* thumbnail of the page (a scaled version of the page as in Fig. 1a). Third, we created an *enhanced* thumbnail, which differed from the plain thumbnail in three ways: (1) the fonts in H1 and H2 tags were modified so that their text would be readable in the thumbnails;[3] (2) highlighted callouts were included for keywords from the search query; and (3) the contrast level in the underlying thumbnail was reduced to enhance the prominence of the callouts (see Fig. 1c).

Creating pages showing collections of summaries. For each of the 12 test questions and 6 practice questions, we chose 100 result pages to present to the participants. We randomized the order of the results as returned by Google. We modified the data set to remove pages that had errors (e.g., the page at the given URL could not be retrieved, or our thumbnail generator did not work for that particular page), to ensure that no answer appeared in the first 10 items of any collection so that the participants would need to examine at least 10 summaries for each question, to minimize the number of duplicates, and so that approximately the same number of correct answers appeared in each question associated with a given category (see Table 1). Finally, we

[2] http://www.google.com/.

[3] Many Web pages do not include H1 and H2 tags, so a large number of pages in our corpus were unaffected by this modification.

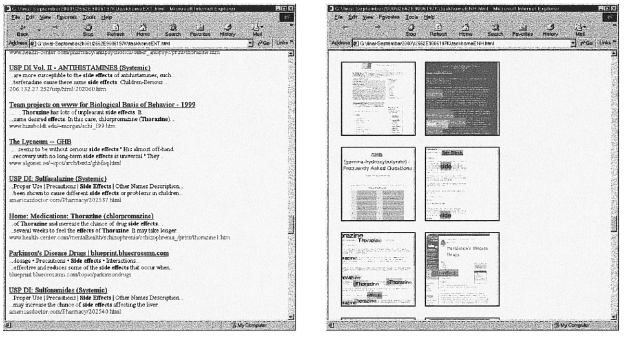

FIG. 2. Browser containing text summaries (left) and browser containing enhanced thumbnails (right).

designed the questions and chose results so that the answer would not be available in any of the summary types (although one might have stronger cues than the other). For example, the Picture questions required sufficient detail that the participant would generally need to visit the actual Web page to be sure they had found a correct answer, and we excluded e-commerce results that explicitly listed the price in the summary. Most summaries did not include such information explicitly, so very few items were excluded on these grounds.

For each question/type of summary (text, plain thumbnail, enhanced thumbnail) combination, we created a single HTML page that contained the summaries of the 100 result pages, with hyperlinks to the actual pages as cached on the local workstation. For the text summaries, the title of the page was a hyperlink. For the thumbnails, the entire thumbnail was a hyperlink.

The text summaries were presented in a single column, using standard Google HTML formatting. The plain and enhanced thumbnails were presented in two columns. We sized the thumbnails to match the size of a typical text summary displayed at a normal font, so as to study the most efficient use of that space: the size of the each thumbnail was 40,000 pixels (200 $\times$ 200), which was approximately the number of pixels occupied by a typical formatted Google result displayed with a standard font size. The vertical spacing between the text summaries was the same size as the vertical and horizontal spacing between the plain and enhanced thumbnails. The browser was a consistent size during all experiments, so that approximately seven text summaries and approximately six thumbnails plus small portions of two additional thumbnails were visible on the screen at a given time (see Fig. 2).

F2

Procedure

After arriving, participants were provided with an overview of the experiment. They were instructed that their task would be to look at collections of links to Web pages and find information contained in those Web pages. The features of Microsoft® Internet Explorer were reviewed, although all participants had at least some familiarity with the browser.

Participants were given the following instructions before beginning the experimental tasks: (1) each link summary page contains 100 links, presented in random order; (2) the only navigable links are those on the link summary page; (3) all searching must occur in a single Microsoft® Internet Explorer window; (4) for some of the questions, there are multiple pages containing the answer; and (5) the answer does not have to be the best one, just the first one that fulfills the requirements of the question.

After finishing the experimental introduction, the participants began the tasks. First, the participant was introduced to one of the three types of summary page, which were presented in a counterbalanced order across participants. The summary page was loaded into the browser window. To the upper left of the browser window was a window containing the question. Below the question window was another window containing Start and Stop buttons. Participants were instructed to press the Start button at the beginning of each search. When they thought they had found the answer, participants pressed the Stop button at which point the experimenter would confirm whether the answer found was acceptable. If the page did not contain an acceptable answer, the participants pressed Start again to continue the search.

Next, participants completed two practice questions to familiarize themselves with that type of summary page. During the practice questions the participants were encouraged to ask for clarification or further instruction as necessary. After finishing the practice questions, the participant completed four test questions using the same type of summary page. Each set of four test questions included one question from each of the four Question categories (i.e., Picture, Homepage, e-commerce, Side-effects). If the participant did not find a correct answer within 5 minutes, they were asked to stop searching and advanced to the next question. When the first set of test questions was finished, the participant repeated the procedure for each of the other two summary page types.

After the participant had answered all questions for all summary types, the experimenter interviewed them about their experiences using the different summary pages. Participants were then thanked and excused. The experiment lasted approximately 75 minutes.

Our instrumentation package consists of a program called WebLogger (Reeder, Pirolli & Card, 2001) that records user gestures (such as keystrokes or scrolling) and actions by the browser application (such as loading and rendering pages). After the experiment was completed, we analyzed the data output by WebLogger to extract timing information and the number of page visits per question. WebLogger also enforced the navigation constraints mentioned above (i.e., WebLogger prevented participants from following links on nonsummary pages).

Results

We analyzed our data in several ways. As a first-order analysis, we looked at how summary type affected total search time and the total number of pages visited. Search can be frustrating to the user both because it may take a long time to find a Web page that answers the user's question, and because in the interim the user will often visit a number of Web pages that do not satisfy the question.

In the second part of our data analysis, we examined the time spent on the *summary* pages and *content* pages (the individual Web pages that potentially contain answers to the questions), as a function of the number of visits to content pages (and correspondingly, revisits to the summary pages) and summary type. The optimal strategy for reducing search time depends in part on where that time is spent. If a user spends a large amount of time on the summary page, then a summary type that reduces the time spent scanning the summary page without sacrificing accuracy will significantly affect total search time. If the user spends a large amount of time analyzing each content page to determine whether or not it satisfies the query, then that suggests the need to help the user accurately select content pages worth visiting.

In general, we adopted a significance level of $p < 0.05$ in our statistical tests. In some cases, we report values of p that approach but do not pass this significance threshold. Such effects that appear marginally significant in the current study might be significant in an experiment with more statistical power (e.g., one that uses more participants). We further investigate the nuances of the data with planned linear contrasts. These linear contrasts compare the text summaries, plain thumbnails, and enhanced thumbnails on subsets of the data (e.g., time to complete the Homepage task). The contrasts, based on a two-tailed t distribution, are conservative tests of the differences between summary types. However, because of the large number of linear contrasts computed, a more strict p-value of 0.01 was used as the significance level for these comparisons.

In the third analysis, we compute a measure, for each of the three summary types, of the *false alarm* rate—the tendency of a user to view a summary and make the incorrect assessment that the corresponding page contains the answer to the user's query. When the false alarm rate is high, much search time can be wasted visiting content pages that do not contain the answer to a question.

Following the quantitative analyses, we briefly review participant responses to the three summary types. Finally, we provide a summary of our findings. In the Discussion section, we integrate these analyses, relating the various metrics to possible search strategies.

Total Search Time and Number of Pages Visited

In this section we present our data on search time and number of pages visited for each summary type. If participants were unable to find an answer in five minutes, their total search time was recorded as 5 minutes.[4] Fn4

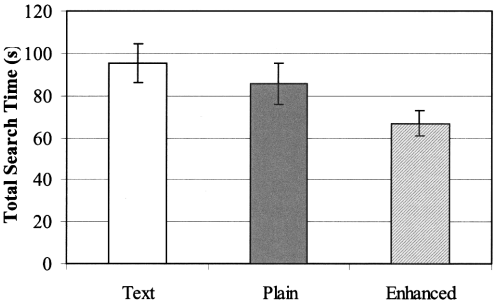

FIG. 3. Total search time for the three summary types across all four question categories. Error bars show the standard error.

Total search time. We performed an ANOVA on total log search time[5] with two within-subjects factors, summary type (text, plain thumbnail, enhanced thumbnail) and question category (Picture, Homepage, e-commerce, Side-effects). Participants needed more time for some question categories than for others; specifically, they were slower to answer Side-effects questions (mean = 126 s, SD = 83.2) than e-commerce (mean = 68 s, SD = 46.6), Homepage (mean = 77 s, SD = 79.6), or Picture (mean = 59 s, SD = 58.1) questions, $F(3,51) = 25.42$, MSE = 0.063, $p < 0.01$. Fn5

Figure 3 shows the total search time for the three different summary types, averaged across all question categories. F3 The time participants needed to answer questions across the three summary types marginally differed, $F(2,34) = 2.75$, MSE = 0.120, $p = 0.08$. To understand the nature of these differences, planned linear contrasts were conducted. Participants answered questions more quickly with enhanced thumbnails (mean = 67 s, SD = 49.9) than with text (mean = 95 s, SD = 78.1; $t(34) = 2.27$, $p = 0.01$), and slightly more quickly with plain thumbnails (mean = 86 s, SD = 84.4) than with text, $t(34) = 1.65$, $p = 0.05$. There were no significant time differences between enhanced and plain thumbnails, $t(34) = 0.62$, $p = 0.27$.

How well participants performed with each summary type varied across the four question categories, Picture, Homepage, e-commerce, and Side-effects, $F(6,102)$

[4] This occurred less than 5% of the time (for 9 of the 216 questions). Seven of the nine times participants failed to find the answer within the time limit, they were trying to answer one of two questions—one a Side-effects question, and one a Homepage question. This suggests that these two questions were more difficult than the other 10 questions. However, participants did not have difficulty answering these questions when provided with enhanced thumbnails: the nine times participants did not find the answer they were using text summaries (4 of 9) or plain thumbnails (5 of 9).

[5] Distributions of data on time to complete a task tend to be log-normal. Throughout this article, we use a log transformation of the search times to ensure that the data are normally distributed, a requirement when performing an ANOVA.

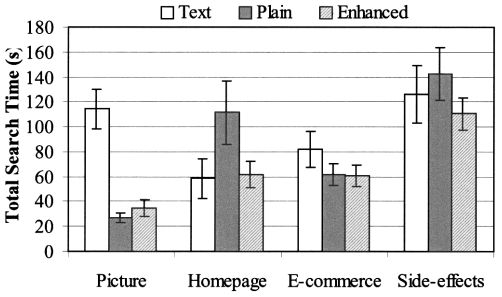

FIG. 4. Total search time for the three summary types, grouped by question category. Error bars show the standard error.

F4

$= 7.97$, MSE $= 0.083$, $p < 0.01$. Figure 4 shows the total search time for the three different summary types and the four question categories. The data are averaged over participants, and within each question category we have averaged over the three questions for that category—thus each bar in the graph represents the average over 18 data points. Separate ANOVAs and linear contrasts were used to compare performance on summary type within each question category.

Summary type had the largest effect on the time needed to answer Picture questions, $F(2,34) = 1.97$, MSE $= 0.062$, $p < 0.01$. In this category, participants were equally fast to answer questions with the plain and enhanced thumbnails, $t(34) = 0.63$, $p = 0.27$, and both thumbnails led to faster performance than text summaries [Plain: $t(34) = 5.25$; Enhanced: $t(34) = 4.62$, $p < 0.01$]. For the Homepage category, we also saw minor differences in summary type, $F(2,34) = 2.72$, MSE $= 0.095$, $p = 0.08$. Time to complete the questions in the text and enhanced thumbnail conditions did not differ, $t(34) = 0.68$, $p = 0.25$, but the text summary times were faster than the plain thumbnail times, $t(34) = 2.04$, $p = 0.03$. For the e-commerce and Side-effects questions, there were no differences in search time across any of the summary types (e-commerce: $F(2,34) = 0.44$, MSE $= .111$, $p = 0.65$; Side-effects: $F(2,34) = 0.38$, MSE $= .101$, $p = 0.69$).

Number of pages visited. We performed an ANOVA on average number of pages visited with two within-subjects factors, summary type (text, plain thumbnail, enhanced thumbnail) and question category (Picture, Homepage, e-commerce, Side-effects). Number of visits did not differ across the question categories [Picture: mean $= 3.9$, SD $= 3.7$; Homepage: mean $= 4.8$, SD $= 6.7$; e-commerce: mean $= 4.4$, SD $= 3.3$; Side-effects: mean $= 5.6$, SD $= 5.0$), $F(3,51) = 1.97$, MSE $= 14.778$, $p = 0.13$]. Number of visits did differ by summary type, with participants visiting fewer pages when they answered questions with enhanced thumbnails (mean $= 3.8$, SD $= 2.9$) or text (mean $= 4.4$, SD $= 3.7$) than when using plain thumbnails (mean $= 5.8$, SD $= 6.9$), $F(2,34) = 2.95$, MSE $= 27.377$, $p = 0.07$. Participants needed fewer visits to answer questions with enhanced thumbnails compared with plain thumbnails, $t(34) = 2.37$, $p < 0.01$, and there was a trend toward fewer visits with text compared with plain thumbnails, $t(34) = 1.64$, $p = 0.06$.

The pattern of page visits for the three summary types varied across the four question categories, $F(6,102) = 7.26$, MSE $= 19.543$, $p < 0.01$. For the e-commerce question, different summary types led to no significant differences in the number of pages visited, $F(2,34) = 0.04$, MSE $= 14.009$, $p = 0.96$. However, visit patterns did differ with summary type for the Picture, $F(2,34) = 28.59$, MSE $= 6.565$, $p < 0.01$, Homepage, $F(2,34) = 4.54$, MSE $= 42.826$, $p = 0.02$, and Side-effects, $F(2,34) = 5.49$, MSE $= 22.607$, $p < 0.01$, questions. Participants answering Picture questions with the plain and enhanced thumbnails needed an equally small number of visits, $t(34) = 0.22$, $p = 0.41$, and those using either form of thumbnail needed fewer visits than those using text summaries [Plain vs. Text: $t(34) = 3.31$; Enhanced vs. Text: $t(34) = 3.09$, $p < 0.01$]. The number of visits needed to answer the Homepage and Side-effects questions were the same for enhanced thumbnails and text [Homepage: $t(34) = 1.21$, $p = 0.12$; Side-effects: $t(34) = 0.54$, $p = 0.30$], and both text summaries and enhanced thumbnails required fewer visits than plain thumbnails [Homepage, Text vs. Plain: $t(34) = 3.70$, Enhanced vs. Plain: $t(34) = 2.48$, $p < 0.01$; and Side-effects, Text vs. Plain: $t(34) = 2.84$, $p < 0.01$, Enhanced vs. Plain: $t(34) = 2.29$, $p = 0.01$].

Summary page time vs. content page time. We next further analyzed search times by splitting them into the time spent on the summary page and time spent on the content pages (the individual Web pages that potentially contain the answers to the questions). Because during the task participants repeatedly switched between the summary page and various content pages, we are particularly interested in the time spent on the summary and content pages per iteration, i.e., per visit to a content page.

We performed an ANOVA on log search time per page visit with three within-subjects factors, summary type (text, plain thumbnail, enhanced thumbnail), question category (Picture, Homepage, e-commerce, Side-effects), and page type (summary page, content page). Overall, participants spent more time on summary pages (mean $= 12$ s/visit, SD $= 8.0$) than on content pages (mean $= 8$ s/visit, SD $= 6.9$), $F(1,17) = 54.84$, MSE $= 0.048$, $p < 0.01$. The interaction between summary type and page type revealed that participants spent the same amount of time per visit on the content pages, but on the summary page enhanced thumbnail users spent less time/visit than plain thumbnail users who spent less time/visit than text summary users, $F(2,34) = 2.74$, MSE $= .031$, $p = 0.08$. The amount of time per visit spent on the summary and content pages varied across question category, $F(3,51) = 30.81$, MSE $= .027$, $p < 0.01$. There was a significant three-way interaction of summary type, question category, and page type, $F(6,102)$

$= 8.21$, MSE $= .022$, $p < 0.01$. Planned linear contrasts were conducted to elucidate these latter two results.

For all question categories except Picture, summary page time per visit was lower for both plain and enhanced thumbnails than for text [Homepage, Plain vs. Text: $t(17) = 8.27$, Enhanced vs. Text: $t(17) = 7.35$, $p < 0.01$; e-commerce, Plain vs. Text: $t(17) = 2.63$, $p < 0.01$, Enhanced vs. Text: $t(17) = 1.88$, $p = 0.04$; Side-effects, Plain vs. Text: $t(17) = 4.55$, Enhanced vs. Text: $t(17) = 3.29$, $p < 0.01$]. For the Picture category, this pattern reversed, with summary page time/visit lower for text than for either plain, $t(17) = 9.06$, $p < 0.01$, or enhanced thumbnails, $t(17) = 8.40$, $p < 0.01$.

The pattern for content page time/visit was somewhat different. For the Homepage category, plain thumbnails led to a marginally lower content page time/visit than text summaries, $t(17) = 1.92$, $p = 0.04$. For the e-commerce category, there was no effect of summary type. For the Side-effects category, content page time/visit was lower for either type of thumbnail than for text [Plain vs. Text: $t(17) = 6.02$; Enhanced vs. Text: $t(17) = 3.83$, $p < 0.01$], and marginally lower for plain than enhanced thumbnails, $t(17) = 2.19$, $p = 0.02$. For the Picture category, text led to lower content page time/visit than either type of thumbnail [Text vs. Plain: $t(17) = 7.43$; Text vs. Enhanced: $t(17) = 6.55$, $p < 0.01$].

View Time and False Alarm Rate Per Summary

Although we did not collect detailed eye movement and mouse movement data, we did collect logs recording the on-screen duration of every summary (e.g., of every plain thumbnail) during the experimental tasks, and the total number of summaries that were viewed (including counts of multiple viewings). From these data we could calculate the average amount of *view time* per summary, for each type of summary. Also, we could calculate the proportion of summaries viewed that resulted in a visit to the corresponding content page. This is a kind of *false alarm* rate—an estimate of the propensity of users to visit links falsely thinking that they are correct. As we will discuss below, small perturbations of this false alarm rate can have dramatic effects on the time costs of surfing hyperlinked content.[6] Fn6

View time per summary. View time per summary is the amount of time that the average summary on the summary page was displayed to a user. For every question answered by a participant we calculated the view times as the total time summaries were displayed to the user divided by the total number of summaries displayed.[7] We performed a repeated measures ANOVA on the log view times, with factors of summary type (text, plain thumbnail, enhanced thumbnail) and question category (Picture, Homepage, e-commerce, Side-effects). The overall mean view time per link was 1.08 s, and this did not differ significantly among the link summary types, $F(2,150) = 0.68$, $p = 0.51$. There was a main effect of question category, $F(3,150) = 11.27$, $p < 0.01$, with the mean view times being Homepage $= 1.06$ s, Side-effects $= 1.06$ s, Picture $= 1.08$ s, and e-commerce $= 1.10$ s. The interaction of summary type with question category was also significant, $F(6,150) = 2.31$, $p = 0.04$. Overall, however, the different types of link summaries do not garner different amounts of view time during user interaction. The magnitude of the view time costs is approximately 1 s per link and the magnitude of even the strong question category differences is on the order of hundreds of milliseconds. This was somewhat to be expected, as the difference in time to skim a text summary as opposed to getting the gist of an image is also on the order of hundreds of milliseconds. Fn7

TABLE 2. False alarm rate estimates.

	Summary type		
Task	Text	Plain	Enhanced
Picture	.0565	.0099	.0143
Homepage	.1350	.0267	.0272
E-commerce	.1359	.0524	.0483
Side-effects	.0413	.0563	.0426

False alarm rate for each type of summary. The number of visits to content pages performed by users showed a linear correlation with the number of summaries the user viewed during a question, $r = .86$. Making a mistaken visit is relatively costly, averaging 7.97 s.

On every question trial, one of the visits was the final correct answer. We, therefore, calculated the false alarm rate as (number of visits $-$ 1)/(total number of links viewed), for every combination of question category and summary type, shown in Table 2.[8] An ANOVA with factors of summary type (text, plain thumbnail, enhanced thumbnail) and question category (Picture, Homepage, e-commerce, Side-effects) yielded no main effect of link summary T2/Fn8

[6] The false alarm rate measures the fraction of cases in which viewing a summary leads users to believe an answer is on a page when it is in fact not present. A corresponding measure, the loss rate, measures the fraction of cases in which viewing a summary leads users to believe an answer is not on a page when it is, in fact, present on that page. In this analysis we consider only the false alarm rate, which is of particular interest to us because of its potentially high impact on search times.

[7] Of the 216 task logs (18 participants $\times$ 12 questions each), we had to eliminate 38 because of log recording problems (less than 18% of the trials). These problem files were distributed randomly across experimental conditions.

[8] Note that one can get a high and unreliable false alarm rate if the number of summaries viewed is low. For example, if a participant finds the correct answer on their second visit to a content page, but has viewed only two summaries, the false alarm rate is high, even though they found the answer very quickly.

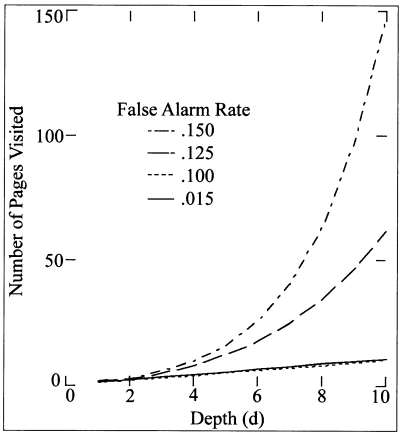

FIG. 5. Effects of perturbations of false alarm rates.

type, $F(2,150) = .36$, $p = 0.70$. There was an effect of question category, $F(3,150) = 15.01$, $p < 0.01$, and an interaction of summary type by question category, $F(6,150) = 5.87$, $p < 0.01$. The false alarm rate for enhanced thumbnails is either the lowest or close to the lowest false alarm rate in each of the question categories.

The false alarm rates vary by a factor of almost 14, from about 0.01 to about 0.14. Although the absolute sizes of these false alarm rates seem small, such variations can have a dramatic impact. This can be illustrated by considering an idealized case of searching for information by surfing along links in a hypertext collection, such as a Web site. Assume that the imaginary Web site is arranged as a tree structure with an average branching factor b. Assume that a user starts at the root page and is seeking a target page that is depth d from the root. If the false alarm rate, f, is perfect, $f = 0$, then the user will visit d pages. This cost grows linearly with d, the distance of the target from the root. If the false alarm rate is maximum, $f = 1$, then the user will visit half the pages in the Web site, on average. This cost grows exponentially with d, because the number of pages grows exponentially with depth.

F5 Fn9 Figure 5 shows the effects of perturbations in false alarm rates more concretely[9] by displaying search cost functions for a hypothetical Web site with branching factor $b = 10$. Search cost refers to the number of pages a user must visit before arriving at the desired page. The curves represents cost functions for links with false alarm rates of $f = 0.015$, 0.100, 0.125, and 0.150. One can see that the search cost regime changes very little as f ranges from 0.015 to 0.100, but changes dramatically as f becomes greater than .100. Indeed, for a branching factor of $b = 10$, there is a phase change from a linear search cost to an exponential search cost at the critical value of $f = 0.100$. Small improvements in the false alarm rates associated with individual links can have dramatic qualitative effects on surfing large hypertext collections.

[9] This search cost analysis follows the analysis of heuristic search developed by Huberman and Hogg (1987).

Participant Response

At the conclusion of the experiment, participants were asked about their search strategies and opinions of the three types of summaries. Several of the participants noted that using the enhanced thumbnails was intuitive and less work than using either the text or plain thumbnails. One participant commented that searching for information with text summaries did not seem difficult before he was exposed to searching with the enhanced thumbnails. Sixteen of the 18 participants used the genre information present in the thumbnails. Fifteen participants used cues from the callouts, specifically the relationship between search terms, the location of search terms, or how often the terms appeared, when searching for information with the enhanced thumbnails. Seven participants rated the enhanced thumbnails as their favorite summary type overall, while an additional six preferred the enhanced thumbnails for certain types of tasks. Those participants who did not prefer the enhanced thumbnails to the plain thumbnails or text summaries reported that they liked the idea, but desired changes in our implementation of the enhancements.

Summary

Here we summarize some of the major results of this analysis:

(1) For total search time, text summaries are the worst overall.
(2) The relationship between summary type and total search time depends greatly on question category.
(3) For minimizing the number of visits to content pages, plain thumbnails are worst.
(4) The relationship between summary type and number of visits depends upon question category.
(5) Participants spent more time on the summary page per visit than on the content pages.
(6) For all but the Picture task, participants spent more time on the summary page per visit with text summaries than with either type of thumbnail.
(7) False alarm rates depended greatly on task, with enhanced thumbnails always yielding either the lowest false alarm rates or nearly the lowest.

Discussion

One of the most interesting results is the fact that the relationship between summary type and total search time is affected so strongly by question category. Here, we examine this result in a more detailed analysis. By considering the results of several of our analyses simultaneously, we see a pattern that suggests that for some question categories, participants used different strategies with one of the summary types than with the others, and that the strategy used for a summary type may vary by question category. We now discuss these possible strategies, relating them to information foraging theory and considering their advantages and

disadvantages. We then review our findings in light of these strategies, showing how they explain user performance in certain tasks. Finally, we discuss design implications.

Our basic conjecture is that searchers use strategies based on cues encountered during the current task. Previous work on information foraging theory (Pirolli & Card, 1999) presents a computational cognitive model to predict *information scent:* the strength of local cues, such as text labels, in providing an indication of the utility or relevance of a navigational path leading to some distal information. The summary types can be considered as having some degree of information scent, i.e., some degree of useful information about whether the page they represent is worth visiting. Participants are likely to use this information to determine whether or not to follow a link. Further, participants are likely to leave a page when they feel that the utility of additional time spent on that page is lower than the utility value of going to another (either summary or content) page. Utility might depend upon both the additional time required on the page and the amount by which this additional time would likely increase the chances of finding the answer to the query.

We now consider two different degrees of information scent. We begin with the case in which the summary type has a high degree of information scent. In this case, participants are likely to use a *high-scent* strategy. In other words, they may use the summary page to fairly carefully identify a summary that is likely to lead to the correct answer. They will visit the corresponding page and search for the answer on this page, repeating the process if the answer, in fact, seems not to be available. Because summaries provide information relevant to the task, we would expect that this case would generally be characterized by a lower number of visits to content pages, lower false alarm rates, and longer visit times on content pages. (Although, of course, one would expect aspects of the content page itself to affect visit times, on average we expect visit times to be longer because (a) the user has already extracted some of the most obvious information about the page from the summary, so the purpose of the visit is to extract more detailed information; (b) the user may be looking for a specific element indicated in the summary; and (c) the user may have higher confidence that the answer is available on the content page, and therefore, be willing to spend more time searching for it. The effect of the content page itself on visit times is beyond the scope of this article.)

Next, we consider the case in which the summary type has a low degree of information scent. In this case, participants are likely to use a *low-scent* strategy: a low amount of information on the summary page might make it worthwhile for a user to choose pages fairly arbitrarily so they can quickly go to content pages that may have better information. Because the summary page provides little information about the content page, a large amount of new information may be quickly and easily available from the content page. The user may quickly extract this information and, if the page does not look promising, return to the summary page. We would expect that this case would generally be characterized by a higher number of visits to content pages, higher false alarm rates, and shorter visit times on content pages.

Although the low-scent strategy may be the best the user can do in a given situation, we believe the low-scent scenario is less desirable than the high-scent scenario. For example, it requires more visits. In a real-world situation, an increased number of visits translates into more time spent on the task because of network latencies in downloading additional content pages (we removed these latencies in our study by caching pages locally). Further, the low-scent scenario is more likely to frustrate participants because it requires more guesswork and gives them less of a sense of control.

Our data suggest that the low-scent strategy was used for text summaries in the Picture category, and for plain thumbnails in the Side-Effects and Homepage categories. We discuss these cases in turn, relating each to our data in the Results section. In the Picture case, participants spent less time on the summary page, less time on the content page, and made more visits to content pages when using text summaries than when using plain or enhanced thumbnails. Further, the false alarm rate was higher for text than the other summary types. Perhaps participants in this situation spent less time on the summary page because for the Picture questions there is more relevant information available on content pages than on the text summary pages. It makes sense that text would be less informative for these questions than thumbnails, as thumbnails allow a user to see the presence of a picture on a page. Once on the content page, in many cases participants may have quickly seen that the top of the page did not contain an image, and judged it more cost-effective to go back to the summary page and try another content page than to further examine the current content page, leading to short visits to content pages. Conversely, the plain and enhanced thumbnail summary pages provided a great deal of information relevant to the Picture questions, so participants may have assessed that their time was well spent on the summary pages. Once they selected a page to visit, they may have had a strong expectation that the correct answer was on that page, and therefore, been willing to spend longer visiting the page.

For the Side-effects category, we see much the same pattern, except with plain thumbnails rather than text summaries leading to the shorter visit times and larger number of visits. Plain thumbnails provide weak information scent relevant to this question category, so users may have quickly made a guess as to what page to visit to proceed to more informative content pages. Once on a content page, participants may have quickly judged that the drug name did not appear in the page header and concluded that they would more efficiently spent their time trying a different content page. Again, such a strategy would lead to relatively more visits and a higher false alarm rate, as found.

In the Homepage questions we see a similar pattern for plain thumbnails compared to text summaries—plain thumbnails lead to a greater number of shorter visits than

text summaries. Again, this behavior makes sense—the name of the person, either in a text summary or an enhanced thumbnail, aids in finding their homepage. Although one can perhaps classify a page as a homepage without this text information, such a classification is sometimes misleading, as search results often include homepages for people other than the target. Participants may quickly visit pages to extract information such as the text in headers. However, in this case plain thumbnails did not yield a higher false alarm rate.

Note that in all cases in which we see evidence of the low-scent strategies, low information on the summary page leads to short visit times on *both* the summary and content pages.

Although we see hints of low-scent strategy in some cases for text and plain thumbnails, we do not see evidence of this strategy for enhanced thumbnails. Instead, we see a pattern in which enhanced thumbnails consistently lead to short visits to the summary page, medium-length visits to the content pages, few visits, and low false alarm rates. This pattern suggests that enhanced thumbnails have high scent, and therefore, consistently allow for quick and accurate judgments about which content pages contain the answer to the query. This apparent *high-scent* effect is particularly interesting because study participants had developed strategies for using text summaries over a period of years and lacked corresponding experience with thumbnails.

This effect translated into benefits for enhanced thumbnails overall. The relative performance of plain thumbnails and text was variable: these two summary types would sometimes yield the best performance (for tasks for which they were particularly well-suited) and sometimes the worst performance (for tasks for which they were a poor fit). Enhanced thumbnails, which combine the features of text summaries and plain thumbnails, were more consistent than either text summaries or plain thumbnails, having for all categories the best performance or performance that was statistically indistinguishable from the best.

High-scent summaries have numerous benefits. Because participants spent more time on summary pages than on content pages, improving the quality of summaries may have a significant impact on the amount of time spent on the task overall. Further, improving the summaries reduces the false alarm rate, and minimizes the chance of the less desirable *low-scent* strategy.

Conclusions and Future Work

We have presented enhanced thumbnails that work to combine the advantages of both text summaries and plain thumbnails. We have conducted a study to compare the performance of enhanced thumbnails with plain thumbnails and text summaries. Across the collection of question categories, we found that enhanced thumbnails yielded the best and most consistent performance.

In addition to conducting further studies, we are pursuing several extensions of this work. We are interested in increasing the information scent of the enhanced thumbnails. For example, we are experimenting with emphasizing in the thumbnails items other than search keywords. One can use a number of algorithms to choose relevant words given a search goal, for example, term frequency inverse document frequency (TFIDF). One can also choose nontextual elements to enhance, for example, by choosing a representative image to enlarge.

We are also exploring different ways to position the callouts on the thumbnail. In the examples presented in this article, the callouts are positioned directly above the word with which they are associated in the thumbnail. However, it may be desirable to slightly adjust the position of the callouts so as to minimize their occlusion of each other or of other useful information on the thumbnail such as readable headers. Another alternative is to include the text in only one callout per thumbnail and render the other callouts as colored bands only, giving the user a sense of the distribution of the word in the page without cluttering it with text.

It would also be interesting to consider how one might build thumbnails into a production search engine. Doing so would introduce many significant system-engineering issues, such as the bandwidth requirements to download the images and the time to generate thumbnails for a given query. Partial precomputation of the thumbnails may address the latter, but would introduce storage requirements.

Acknowledgments

We are very grateful to Rob Reeder for providing assistance with WebLogger, Pam Schraedley for her contributions to the statistical analysis, and Paul Aoki for helpful discussions and comments. This research was funded in part by Office of Naval Research Contract No. N00014-96-C-0097 to Peter Pirolli and Stuart K. Card.

References

Ayers, E., & Stasko, J. (1995). Using graphic history in browsing the World Wide Web. Proceedings of the 4th International World Wide Web Conference.

Card, S.K., Robertson, G.G., & York, W. (1996). The WebBook and the Web Forager: An information workspace for the World-Wide Web. Proceedings of CHI'96, ACM Conference on Human Factors in Computing Systems (pp. 111–117).

Chapman, A. (Ed.). (1993). Making sense: Teaching critical reading across the curriculum. New York: The College Board.

Cockburn, A., Greenberg, S., McKenzie, B., Jasonsmith, M., & Kaasten, S. (1999). WebView: A graphical aid for revisiting web pages. Proceedings of OZCHI'99, Australian Conference on Human Computer Interaction.

Coltheart, V. (Ed.). (1999). Fleeting memories: Cognition of brief visual stimuli. Cambridge, MA: MIT Press.

Czerwinski, M.P., van Dantzich, M., Robertson, G., & Hoffman, H. (1999). The contribution of thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3D. Proceedings of INTERACT'99, 7th IFIP Conference on Human-Computer Interaction (pp. 163–170).

Hearst, M. (1995). TileBars: Visualization of term distribution information in full text information access. Proceedings of CHI'95, ACM Conference on Human Factors in Computing Systems (pp. 59–66).

Helfman, J.I. (1999). Mandala: An architecture for using images to access and organize web information. Proceedings of Visual '99, Visual Information and Information Systems, Third International Conference (pp. 163–170).

Hightower, R., Ring, L., Helfman, J., Bederson, B., & Hollan, J. (1998). Graphical multiscale web histories: A study of PadPrints. Proceedings of Hypertext '98 (pp. 58–65).

Hoffman, J.E. & Mueller, S. (1994, November). An in depth look at visual attention. Paper presented at the 35th annual meeting of the Psychonomic Society, St. Louis, MO.

Huberman, B.A., & Hogg, T. (1987). Phase transitions in artificial intelligence systems. Artificial Intelligence, 33, 155–171.

Jansen, B.J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. SIGIR Forum, 32(1), 5–17.

Kopetzky, T., & Mühlhäuser, M. (1999). Visual preview for link traversal on the WWW. Proceedings of the 8th International World Wide Web Conference (pp. 447–454).

Lankheet, M.J.M., & Verstraten, F.A.J. (1995). Attentional modulation of adaptation to two-component transparent motion. Vision Research 35, 1401–1412.

Lorch, R.F. & Myers, J.L. (1990). Regression analyses of repeated measures data in cognitive research. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 149–157.

Morrison, J.B., Pirolli, P., & Card, S.K. (2001). A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. Extended Abstracts of CHI'01, ACM Conference on Human Factors in Computing Systems (pp. 163–164).

Pirolli, P., & Card, S.K. (1999). Information foraging. Psychological Review, 106, 643–675.

Paivio, A. (1974). Pictures and words in visual search. Memory & Cognition, 2(3), 515–521.

Reeder, R.W., Pirolli, P., & Card, S.K. (2001). Web EyeMapper and WebLogger: Tools for analyzing eye tracking data collected in Web-use studies. Extended Abstracts of CHI'01, ACM Conference on Human Factors in Computing Systems (pp. 19–20).

Robertson, G., Czerwinski, M., Larson, K., Robbins, D.C., Thiel, D., & van Dantzich, M. (1998). Data mountain: Using spatial memory for document management. Proceedings of the ACM Symposium on User Interface Software and Technology '98 (pp. 153–162).

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. Vision Research, 39, 3157–3163.

Technology Review. (2000). Search us, says google. Technology Review. http://www.technologyreview.com/magazine/nov00/qa.asp.

Wind River. (2000). ICE Browser 5. http://www.icesoft.no/icebrowser5/index.html.

Winer, B.J. (1971). Statistical principles in experimental design. New York: McGraw-Hill.

World Wide Web Consortium. (1998a). Cascading style sheets, level 2 (CSS2) specification. http://www.w3.org/TR/REC-CSS2/.

World Wide Web Consortium. (1998b). Document object model (DOM) level 1 specification, version 1.0. http://www.w3.org/TR/REC-DOM-Level-1/.

Woodruff, A., Faulring, A., Rosenholtz, R., Morrison, J., & Pirolli, P. (2001). Using thumbnails to search the Web. Proceedings of CHI'01, ACM Conference on Human Factors in Computing Systems (pp. 198–205).

Wynblatt, M., & Benson, D. (1998). Web page caricatures: Multimedia summaries for WWW documents. Proceedings of the IEEE International Conference on Multimedia Computing and Systems (pp. 194–199).

AQ: 1

AQ: 2