

ENHANCING THE READABILITY OF SEARCH RESULT SUMMARIES

Anne Aula

Tampere Unit for Computer Human Interaction

Information Visualization Research Group

Department of Computer Sciences, University of Tampere, Finland

E-mail: anne.aula@cs.uta.fi

ABSTRACT

Most current web search engines use similar textual search result lists. Their efficiency is largely untested, as are the alternatives for facilitating the readability of the textual summaries. This paper presents an experiment on the efficiency of three textual result summary styles: a novel summary style utilizing bulleted lists (*list*), a normal Google-like style with search term bolding (*normal-bolded*), and the same style without search term bolding (*normal-plain*). 27 participants completed ten question-answering tasks with each summary style. The novel list style was observed to provide clear benefits as the task times were almost 20 % faster with list summaries than with normal-bolded summaries. The search term bolding had very little or even negative effect on task times. In sum, the readability of textual result summaries can be facilitated by simply re-organizing their layout.

Keywords

Search engines, result summaries, readability.

1. INTRODUCTION

Most web search engines show search results as a textual list. Although some visualizations of the results are already available in World Wide Web (web), for example, Kartoo (www.kartoo.com) and WebBrain (www.webbrain.com) or proposed in the literature, for example, TileBars [2] and Envision [7] (for an overview, see [17]), textual results are still the most commonly used result presentation style. One important benefit of textual result lists over visualizations is that they do not require the users to learn new ways of interacting with them – they are readily understandable even for novice users [11, 17].

Textual search results have a document summary that is typically *query-biased* [13], meaning that it only has

sentences that contain the user's search terms and not, for example, the first couple of sentences from the document. There are two main variations of the common result style: the user's search terms are either shown in bold or not. Term bolding is used by, for example, AltaVista (www.altavista.com), Google (www.google.com), and Yahoo (www.yahoo.com). The trend seems to be towards bolding the search terms, but, for example, Kartoo's HTML version (www.kartoo.com/en/kartoo.html) does not bold them (as of July 2004).

Surprisingly, the efficiency of the common textual result styles is largely untested. An exception is a recent study suggesting that even random bolding of words in the summary improves the accuracy of relevance evaluations [3]. The possible effects of search term bolding can also be approached theoretically. Visual search for salient objects is very fast; distinct objects (*e.g.*, bolded terms in text) pop-out from other objects already pre-attentively [1, 4, 14, 15, 16]. Thus, when search terms are bolded, the user does not have to engage into a slow serial search since the terms will clearly stand out from the other text. From the bolded terms, the user can easily make sense of the textual context in which the terms appear. However, the readability of the text with bolded words is another issue as some highlighting techniques (such as underlining) distract reading [8]. Although search term bolding is widely used and there is some evidence that it is beneficial for search result evaluation [3], the question of whether it affects the readability of the summaries requires further study.

A textual result page is a list of individual results. Each result can further be divided into smaller entities of information (title, summary, and URL). Usually the different parts of the result are rendered in different font or colour, which makes them perceptually distinguishable from one another and facilitates information processing [5, pages 89-91]. The summaries again consist of smaller entities of information, namely, excerpts of different sentences. In query-biased summaries, the excerpts are from different parts of the document, so the text does not proceed logically from the beginning of the summary to the end. To show this to the user, the different sentences are typically separated from one another with an ellipsis ('...'). However, the ellipsis might not be an optimal solution: the sentences in the summary are essentially a list of sentences

List:

HCI 2004 Design for Life: Upcoming Deadline: 7th May 2004...

- Annual Conference is taking place at **Leeds** Metropolitan University...
- May 7th **2004** is the deadline for industry...
- ...concerns are traditional ones for **HCI**; others are...
www.chiplace.org/modules.php?op=modload&name=News&file=article&sid=237

Normal-bolded:

HCI 2004 Design for Life: Upcoming Deadline: 7th May 2004 ...

... Annual Conference is taking place at **Leeds** Metropolitan University ... May 7th **2004** is the deadline for industry ... concerns are traditional ones for **HCI**; others are ...
www.chiplace.org/modules.php?op=modload&name=News&file=article&sid=237

Normal-plain:

HCI 2004 Design for Life: Upcoming Deadline: 7th May 2004 ...

... Annual Conference is taking place at Leeds Metropolitan University ... May 7th 2004 is the deadline for industry ... concerns are traditional ones for HCI; others are ...
www.chiplace.org/modules.php?op=modload&name=News&file=article&sid=237

Figure 1. Examples of the summary styles studied in the experiment (query: *hci 2004 leeds*)

derived from the document and the ellipsis does not reveal that. Secondly, the ellipsis has different meanings in the summaries: It indicates that the sentence is not complete, that the sentences do not appear together in the underlying document, or that the document continues after the last sentence chosen for the summary. Assigning different meanings to the same visual symbol breaks a well-known usability heuristic of consistency [6] and presumably makes it more difficult to scan the results efficiently.

Our solution to these problems is to place the different sentences in a bulleted list. This procedure enhances the visual regularity of the result page and by that, allows users to “scan ahead” to relevant information [5, page 42] and skip irrelevant sentences quickly by moving their eyes to the next list item. In addition, textual context is known to affect the reading process and the understanding of written material so that if the context is not logical, the processing of text slows down (see [9] for a review). Thus, we wanted to separate the different logical entities (sentences) clearly from one another. Furthermore, to avoid assigning different meanings to the same visual object, we used the ellipsis only for indicating the incompleteness of the sentence.

This study focused on the readability of three different summary styles (Figure 1). The research questions were:

1. Which of the current summary styles (normal-bolded or normal-plain) is more efficient to process?
2. Do the novel list summaries improve the readability of the summaries over the style used by Google (currently the most popular web search engine [10])?

To answer these questions, an experiment was conducted.

2. EXPERIMENT – METHODS

In experimental studies, the confounding variables need to be carefully controlled. We studied the readability of result lists. For this purpose, it would be inappropriate to have participants formulate their own queries as it would result in each of them getting different results. Instead, a controlled setup in which all users see the same result pages ensures that the observed differences in performance are due to the designed modifications.

2.1 Participants

Initially, 27 students and staff members from the University of Tampere participated. Due to technical problems, data from 5 participants was discarded. Additionally, data from 2 participants was discarded because they had over 50 % error rate in the task. To keep the design balanced, 7 new participants were recruited and tested in the corresponding positions in the experimental design. Finally, there were 27 participants (17 males and 10 females) whose average age was 27.0 years (from 19 to 44 years). The participants were experienced in using both computers and web (mean 11.1 and 6.9 years of active usage, respectively).

2.2 Apparatus and Materials

The experiments were run on a Pentium III computer with a 17 inch monitor on a 1024 × 768 resolution. Internet Explorer 6.0 was used as the web browser.

We generated 30 question-answering tasks and associated queries (e.g., Who is the president of the University of X?)

→ query: principal university of x). The average number of query terms was 2.3, which is similar to the typical query length in web search (2.4 in [12]). These pre-determined queries were submitted to Google. From Google’s result list, the first result containing the answer and the first nine results not containing the answer were saved. If nine distracters and one hit could not be found from the first 20 results, another query was formulated. Therefore, too accurate and inaccurate queries were considered to be inappropriate. The 30 tasks were divided into three sets (10 tasks each). In each set of ten tasks, the correct answer to the question appeared exactly once in each position of the list.

All 30 result lists were modified to each of the three result presentation styles (shown in Figure 1):

- *Normal-bolded* style: similar results as in Google’s default style.
- *Normal-plain* style: similar results than in normal-bolded style, except that all bolding was removed.
- *List* style: Google’s summaries were changed so that every time an ellipsis appeared in the summary, a new list item was created. Each list item was preceded with an arrow. An ellipsis was put in the beginning of the sentence if it began with a lower-case letter or a number. An ellipsis was put in the end of the sentence if it did not end with a dot, exclamation mark, or a question mark. The ellipsis was not bolded as it is in Google’s style and the extra space between the ellipsis and the sentence was removed.

All result pages were simplified so that they only had the query field, the Search button, and the result list.

2.3 Procedure

The participants were given the task instructions on paper. They were told that they should try to find the correct answer from the result list and click on its title as quickly as possible, but avoid making mistakes. In the experiment, the participant was presented with a page containing the question. On this page, participants also indicated whether they knew the answer to the question and typed the answer in a textbox if needed. After this, the participants proceeded to the result listing by clicking a “Start task” button. In the result page, clicking the title of the result took the participant to the next task page. At the end of the experiment, the participants submitted their background information (e.g., age, computer and web expertise) using a web questionnaire.

2.4 Design

The experiment used a within subjects design (all participants completed all 30 tasks, 10 tasks with each summary style). The participants were divided into nine groups so that all task set and summary style combinations (3 × 3) occurred equally often. Inside the three task sets, the order of the tasks was randomized. A Latin Square

design was used for controlling the possible effects of summary style and task set presentation orders.

3. RESULTS

Repeated measures analysis of variance (ANOVA) and Bonferroni corrected t-tests were used for task times and error rates. For subjective ratings, Wilcoxon matched-pairs signed-ranks test was used.

Overall, 21 % of the selections were erroneous (did not contain an answer to the task). There were no significant differences between the error rates with different styles and because of that, the errors are left to the task time analysis. An overview of the task time results is given in Figure 2.

	Normal-plain	Normal-bolded		List
Average task times	23.5	25.6	*	20.5
Median task times	19.8	20.6	*	17.2
Average when the answer was not known	23.3	* 26.8	*	21.0

Figure 2. Summary of task times (in seconds). The asterisk denotes a statistically significant difference.

ANOVA showed a significant effect of summary type on task times, $F(2, 52) = 8.78, p < .001$. Pairwise comparisons showed that the task times were shorter with list than with normal-bolded summaries, $t(26) = -4.12, p < .01$. The difference between normal-bolded and normal-plain was not significant, $t(26) = 2.05, p = ns$. In order to ensure that outliers did not cause the observed differences, we also analysed the timing data with median times (with Wilcoxon matched-pairs signed-ranks test). This analysis confirmed that the differences between the summary styles were not resulting from outliers.

In 8.8 % of the tasks, the participants knew the correct answer beforehand, which is not the case in real-life question-answering. Therefore, we analysed separately the cases in which the participant did not know the answer in advance. Again, ANOVA showed a significant effect of summary type on task times, $F(2, 52) = 10.51, p < .001$. Pairwise tests showed that the task times were shorter with list than with normal-bolded summaries, $t(26) = -4.76, p < .01$. The difference between normal-plain and normal-bolded summaries was also significant, $t(26) = -3.09, p < .01$, the task times being slower with normal-bolded.

4. DISCUSSION

This study focused on two questions: First, whether the query term bolding affects the readability of search result

summaries and second, whether the processing of the textual result summaries can be made more efficient by re-organizing their contents. The results showed that the list style significantly improved performance in the experimental tasks over the normal-bolded style (Google-like). Query term bolding had very little or even negative effect on performance.

In this study, the task was to find answers from the result listing and not to use summaries for evaluating the relevance of the underlying document. However, the faster task times with list summaries indicated that the processing of the textual information was facilitated when the different excerpts were organized in a list. The increased efficiency is expected to help the users whenever they need to process the information given in the summaries – regardless of the goal of the search. In exploratory search tasks, the user needs to decide whether the underlying document is related to the search topic at hand. In that case as well, the user needs to process the textual information at some level and list summaries are expected to make this more efficient. We are currently running experiments to verify this.

Query term bolding seemed to slow down the performance in the current study. However, we expect that the type of the search task (exploratory vs. question-answering) affects the usefulness of the query term bolding markedly. In our previous (unpublished) studies, the searchers commented that the distribution of the bolded query terms affects their primary relevance decisions: if the terms are together in the summary, the summary is more likely to be relevant. The actual reading of the summaries is likely to happen only after this initial selection. Previous research has also found some support for term bolding [3], so more research is clearly needed before definite conclusions can be drawn about the benefits of term bolding in result summaries.

When considering the wide-spread use of search engines (250 million searches were done with Google daily in February 2003 [10]), almost 20 % faster task times with list summaries compared to that of the Google-like style is a marked improvement that could save considerable amounts of time for millions of people. The fact that the list style can be automatically applied makes it a promising method for improving the efficiency of textual result summaries.

5. ACKNOWLEDGMENTS

I would like to thank Professor Kari-Jouko Räihä and Natalie Jhaveri for valuable comments, and the participants for their time. The study was supported by the Graduate School in User-Centered Information Technology and the Academy of Finland (project 178099).

6. REFERENCES

- [1] Halverson, T. and Hornof, A.J. (2004). Link colors guide a search. *Proc. CHI 2004*, ACM Press, 1367-1370.
- [2] Hearst, M. (1995). TileBars: visualization of term distribution information in full text information access. *Proc. CHI 1995*, ACM Press, 59-66.
- [3] Kickmeier, M.D. and Albert, D. (2003). The effects of scanability on information search: An online experiment. *Proc. Volume 2 of the HCI 2003: Designing for Society*, Research Press Int., 33-36.
- [4] Matlin, M. (2002) *Cognition*. New York: Thomson Learning.
- [5] Mullet, K. and Sano, D. (1995). *Designing visual interfaces: Communication oriented techniques*. California: Prentice Hall.
- [6] Nielsen, J. (1993). *Usability Engineering*. San Diego, CA: Academic Press.
- [7] Nowell, L.T., France, R.K., Hix, D., Heath, L.S. and Fox, E. (1996). Visualizing search results: some alternatives to query document similarity. *Proc. SIGIR '96*, ACM Press, 67-75.
- [8] Obendorf, H. and Weinreich, H. (2003) Comparing link marker visualization techniques – Changes in reading behaviour. *Proc. WWW 2003*, ACM Press, 736-745.
- [9] Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 3, 372-422.
- [10] SearchEngineWatch – Searches per day. <http://searchenginewatch.com/reports/article.php/2156461>.
- [11] Sebrechts, M.M., Vasilakis, J., Miller, M.S., Cugini, J.V. and Laskowski, S.J. (1999). Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. *Proc. SIGIR '99*, ACM Press, 3-10.
- [12] Spink, A., Wolfram, D., Jansen, M.B.J., and Sarasevic, T. (2000). Searching the Web: The Public and Their Queries. *Journal of the American Society for Information Science and Technology*, 52, 3, 226-234.
- [13] Tombros, A. and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. *Proc. SIGIR '98*, ACM Press, 2-10.
- [14] Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.
- [15] Ware, C. (2000). *Information visualization: Perception for design*. San Diego: Academic Press.
- [16] Ware, C. (2003). Design as applied perception. In Carrol, J. (Ed.). *HCI models, theories, and frameworks: Towards a multidisciplinary science*, 11-26. Morgan Kaufmann.
- [17] Zamir, O. (2001). Visualization of Search Results in Document Retrieval Systems. *SIGTR Bulletin*, 7, 2.

