Searching in a book

Veronica Liesaputra, Ian H. Witten and David Bainbridge

Department of Computer Science, University of Waikato
Hamilton, New Zealand
{vl6, ihw, davidb}@cs.waikato.ac.nz

Abstract. Information has no value unless it is accessible. With physical books, most people rely on the table of contents and subject index to find what they want. But what if they are reading a book in a digital library and have access to a full-text search tool?

The paper describes a search interface to Realistic Books, and investigates the influence of document format and search result presentation on information finding. We compare searching in Realistic Books with searching in HTML and PDF files, and with physical books.

Keywords: Within-Document search, Electronic book, Flash application.

1. Introduction

Libraries and others are digitizing books and enabling people around the world to access them online. These projects enable full-text retrieval over vast collections of books. But once a book has been reached, how do users find information inside and navigate within it?

Subject indexes are an important access tool for physical books, whose utility frequently depends on the quality of the index. As large volumes of text migrate to electronic formats, it is easy to assume that such indexes become superfluous.

Most electronic document readers have a text search function. Readers type what they seek into the search box and are taken to the next closest occurrence, with matching strings highlighted and buttons to move to the next and previous occurrences. However, literal string matching has limitations. Readers who do not know exactly what they seek or how it might be expressed may issue search queries that match many irrelevant passages—or no passages at all.

This paper examines how people retrieve relevant and appropriate passages from within a document. We describe a user study comparing search performance with the Realistic Book format [1] against HTML, PDF and physical books, based on outcome and process measures [3].

2. Experimental Procedure

We conducted a user study that measured performance with each format—HTML, PDF, Realistic Books, and physical books—in terms of both outcome and process measures [3]. Outcome measures concern what readers get from the text. They evaluate *efficiency*, the time spend by readers in answering each question, *effectiveness*, the number of questions answered correctly, and *searching experience*, a subjective assessment of usefulness, likea-

bility, and ease of use. Process measures investigate what search and navigation functions participants use to complete the task. The experiment is designed to test four hypotheses:

H1: Realistic Book users will find answers faster than with other document formats.
H2: Realistic Book users will make fewer task errors than with other formats.
H3: Realistic Book users will report higher satisfaction than with other formats.
H4: Regardless of format, users will use the most effective search tools to answer questions, i.e. BoB for index questions and search tools for full-text search questions.

To help mitigate variability, we recruited 32 high school and university students aged 15–40 from a variety of disciplines. In order to bias the experiment *against* the Realistic Book format, participants were already familiar with the HTML and PDF formats.

For our experiments we used a book that was available in electronic form and had a professional-quality subject index whose terms accurately represented the contents. We chose a university-level text called *Data Mining: Practical Machine Learning Tools and Techniques* [4]. We used only the first three chapters, and modified the table of contents and subject index to remove all references to other parts of the book, retaining the original design and layout.

The book was presented in three electronic formats: HTML, PDF and Realistic Books. Each contained the title, a hyperlinked table of contents (ToC), the main text, and a hyperlinked back-of-book subject index (BoB). To eliminate layout effects all three formats were paginated in exactly the same way as the physical book.

There were four sets of tasks, which participants undertook in the same order. Each task asked four questions that involved seeking information in the book. A different format was used for each task, and participants were exposed to the formats in different orders. The 24 possible orderings were allocated evenly: eight participants performed each task using the same document representation. Because participants did not know they were being timed, they felt no pressure and worked at their own pace. The functions they used were recorded.

To address H4, each task posed two full-text search questions and two BoB questions, in an order that changed from one task to another. Thus participants would use search tools or the BoB at least once while searching with each document format. To provide a balanced assessment of the utility of search, the questions have varying degrees of difficulty: one easy and one hard question for each question type.

3. Results

3.1. Efficiency

A one-way analysis of variance shows that differences due to format are statistically significant at the 1% level for all tasks except the first. For the last 3 tasks, the improvement from Realistic to Physical books is statistically significant at the 1% level (t-value ranges from 3.8 to 4.3), and its improvement over PDF and HTML is statistically significant at the 5% level (t-value from 2.1 to 2.6). The differences between the Physical book and the PDF and HTML formats are significant at the 10% level (t-value from 1.7 to 2.0). There is no significant difference between PDF and HTML for any task.

H1 (speed): Participants consistently produce answers quicker with the Realistic Book.

3.2. Effectiveness

Participants relied on full-text search to find answers in the electronic document formats. Because the subject index questions were designed to be difficult to answer without using the BoB, participants generally got them wrong. With the physical book, most participants failed on the full-text search questions.

H2 (accuracy): A one-way analysis of variance found no significant differences between document formats in the number of task errors made. Thus H2 is not upheld.

3.3. Search experience

Having completed all tasks, participants were asked which formats were most useful, easy to navigate and locate information, pleasant and engaging, and preferred overall.

H3 (satisfaction): Participants found Realistic Books the most useful, engaging and easy to use format, combining a good reading environment with a good searching experience.

The formats were judged useful for different types of documents and activities. HTML is preferred for short documents. Physical books are best for reading activities, but not for information seeking. For searching, PDF or Realistic Books are preferred.

Participants found it hard to understand the structure of the document in HTML and PDF. They easily became disoriented, not knowing where they were in the document and finding it difficult to return to a specific location. Some participants (20%) found it difficult to step through the search results in PDF. Participants felt that they always knew where they were in Realistic Books, and navigated around more freely than with other formats. Book-mark tabs and page edges make it easier to move between search results without losing orientation.

3.4. Participants' strategies

Before beginning each task, most participants briefly overviewed the document and went to the ToC page.

In physical book, they then sought terms they thought relevant in the subject index. If they failed to find the answer through the ToC and BoB, they carefully read sections of the book they thought might contain the relevant information.

In electronic formats, all participants typed into the Find box a word or a phrase that they thought would lead to the relevant information, to see whether it returned any results.

In HTML and PDF, the only time most participants used the BoB was when a search term appeared in it and nowhere else. Nearly all of them neglected the BoB, even though they had been told about it before beginning their tasks and it is listed in the book's ToC, relying instead on search for navigation.

In the Realistic Book, they also used BoB to seek more appropriate search terms when full-text search failed to find the answer, or when they could not think of any more suitable search terms. Once participants were familiar with the document structure, most used the spatial reference of passages in the book to help them remember where information was.

Having exhausted all search terms they could think of, participants would go to the ToC and guess which section contains the information they seek.

H4 (appropriate choice of search tool): Full-text search is the principal information-finding strategy for all formats; participants consulted the BoB as a secondary strategy only for the Realistic Book format. Thus H4 is partially upheld.

4. Conclusions

Back-of-book indexes are considered by readers to be important access mechanisms for physical books. Index terms are carefully chosen to represent the key ideas in the book, and to bridge the author's perspective of the topic to keywords that readers might use. A comparative evaluation of a subject index and full-text search showed that participants found information more effectively and efficiently with the former [2]. However, a post-task questionnaire revealed that participants still preferred search to the subject index.

This paper investigates whether document format influences search behavior and performance. We observed participants performing information-finding tasks in HTML, PDF, and Realistic Book formats, and in physical books. The tasks were designed to test four specific hypotheses.

The results upheld Hypothesis 1: users of Realistic Book found answers significantly faster (99% confidence level) than with conventional document displays.

Hypothesis 2, that Realistic Book users make fewer task errors than with other formats, was not upheld. There were no significant differences between Realistic Book and other formats in terms of the number of correctly answered questions.

Hypothesis 3, that Realistic Book users report higher satisfaction than with other formats, was strongly upheld (although no statistical analysis was performed).

As for Hypothesis 4, that regardless of format users choose the most effective search tools to answer questions, we found that all participants chose full-text search first, whether it was appropriate or not. Only with Realistic Books did they consult the back-of-book index as a secondary choice; for the HTML and PDF formats nearly all participants neglected it. Thus Hypothesis 4 is partially upheld; in addition, we found that whether the back-of-book index is used at all depends on the document display format.

The only time at participants used the back-of-book index for HTML and PDF documents was when one of their search results matched a term in it. Such terms were usually synonyms or bridging words that point users to the actual word used in the text. Without them, readers would not have been able to find the answers. In contrast, readers of Realistic Books also used the back-of-book index to suggest more appropriate search terms when their initial full-text search failed to find an answer, or when they could not think of any more suitable search terms. Thus it is helpful for electronic documents to have well-constructed subject indexes.

Overall, the results reported here strongly uphold the idea that Realistic Books are an effective document format for finding information compared with other formats.

Acknowledgments. We acknowledge the entire New Zealand Digital Library Project team for their unstinting work in providing an environment that makes this kind of research meaningful—and fun. We also acknowledge the support of the European Media Lab where part of the work was done. This research is funded in part by Google.

5. References

1. Liesaputra, V., Witten, I.H. and Bainbridge, D.: Creating and reading Realistic Electronic Books. Computer, **42**(2): 46–55, February/March (2009)
2. Ryan, C. and Henselmeier, S.: Usability testing at Macmillan. Keywords, **8**: 188–202 (2000)
3. Schumacher, G. and Waller, R.: Testing design alternatives: A comparison of procedures. Designing usable texts. Orlando, FL (1985)
4. Witten, I.H. and Frank, E.: Data mining. San Francisco, CA (2005)