

How Should Users Access the Content of Digital Books?

Nina Wacholder
Rutgers University SCILS
4 Huntington
New Brunswick, NJ 08901
732-932-7500 x8214
nina@scils.rutgers.edu

ABSTRACT

I report briefly on some of my own work in each of these areas and elucidate some of the questions that this research has raised. Then I propose as a research agenda the development of a digital library environment containing a suite of inter-related tools specifically designed to facilitate non-sequential access to portions of full-text books and other relatively long documents.

Categories and Subject Descriptors

H.3.7 Information Systems, INFORMATION STORAGE AND RETRIEVAL, Digital Libraries

General Terms

Human Factors

Keywords

Electronic books

Michael Gorman, former president of the American Library Association, declared that “massive databases of digitized whole books, especially scholarly books, are expensive exercises in futility.” (Gorman 2004). His criticism was based on the claim that (scholarly) non-reference books must be read sequentially, from end to end. “A snippet from Page 142 must be understood in the light of pages 1 through 141 or the text was now worth writing and publishing in the first place.” Gorman’s glib refusal to acknowledge any value in scholarly books in digital format denies the legitimacy of reading only part of a book, thereby masking the interesting questions that we can and should ask about how people access the content of electronic books. Among the issues we should study are: 1) the book selection process; 2) content-browsing tools (e.g., search interfaces; tables-of-contents and indexes); 3) different kinds of user needs, as related to individual and task differences; and 4) index term quality. Research on these issues will advance our understanding of how readers access book content and thereby promote the design of more usable and useful electronic books.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

BooksOnline '08, October 30, 2008, Napa Valley, California, USA.
Copyright 2008 ACM 978-1-60558-249-8/08/10...\$5.00.

In what follows, I report briefly on some of my own work in each of these areas and elucidate some of the questions that this research has raised. Then I propose as a research agenda the development of a digital library environment containing a suite of inter-related tools specifically designed to facilitate access to portions of full-text books and other relatively long documents.

1. Some research results and related questions

- **The book selection process** begins when an individual has an information need that may be satisfied by a book (whether print or electronic) and ends with the user’s decision as to whether the book contains enough information on the topic at hand that it merits more in-depth reading. During this process, the user engages in a combination of information-seeking, navigation and reading to make a decision about the book’s usefulness for the current task. To make the decision that a book is useful, the user has to find evidence of some relatively substantial amount of relevant content. To make the decision that a book is not useful, the user has to get enough of an overview of the book to be confident that the book does not have useful content.

Wacholder, Liu and Liu (2006) conducted a study whose objectives were to compare effectiveness of book selection with print and electronic books. Our results showed, unsurprisingly, that overall users did equally well with selection of print and digital books. I do not particularly recommend that researchers continue to compare book selection of print and electronic books, but there are quite a lot of interesting questions that could be asked about what actions people take to make the decision as to whether or not a non-fiction book is useful for a particular topic. How much of the book does a person look at? How accurate are the decisions about usefulness of books for a given topic? What are the barriers to accurate and effective book selection? In what ways do different kinds of information access tools help or hinder the book selection process? How useful are different kinds of metadata (e.g., professionally assigned subject headings or user-assigned tags) in helping or hindering book selection?

- **Content browsing tools** help users explore the content of a book in contexts where reading the whole book may not be appropriate. These tools are useful when a user is interested in something less than the full text of the book. This might happen during the book selection process, or when the user is interested in exploring sub-themes.

The traditional access tools for print books are tables-of-contents and back-of-the-book indexes; with print books

users may also flip through the pages to see if something of interest catches their eye. In the electronic environment, full-text searching is the norm. The enormous usefulness of full-text searching is not in dispute. Still, it is short-sighted to assume that tables-of-contents and indexes are rendered useless by the availability of full-text searching. A more promising approach is to ask whether full-text searching by itself provides adequate access to book content. Should full-text searching be supplemented by tools such as hyper-linked indexes and tables-of-contents? Are the types of indexes and tables-of-contents that are used for print books equally appropriate for digital books? Assuming that they are useful, how specific or exhaustive should the table-of-contents or the terms in the index be? How can the usefulness of these tools be optimized?

- **Index term quality.** In information science terminology, back-of-the-book indexes are *displayed*, i.e., they can be browsed by the reader to get an understanding of what topics are discussed in the books and where in the book the discussion is located. Displayed indexes can be contrasted with non-displayed indexes, such as those used in information retrieval systems and search engines. Many important information retrieval studies have suggested that index terms do not have a crucial impact on search outcome (e.g., Sparck Jones (2005); Salton (1986); Keen (1973)). However, these observations about index term quality apply only to non-displayed indexes.

As noted above, the availability of full-text searching may or may not mitigate the need for displayed indexes for full-text books. These indexes will be most useful when a user does not know what term to enter into the search engine. And if displayed indexes are needed (and I personally believe that they are), then it is important to ask what algorithms or methods produce optimal index terms.

Wacholder and Liu (2006) and Wacholder and Liu (2008) report on a study in which participants used index terms identified by different methods to find answers to questions in a 350 page book (Rice et al. 2001). The results showed that a) users preferred index terms that were longer and more specific; and 2) two sets of index terms (one identified by a program with some degree of linguistic sophistication and another based on the index included in the original print version of the book) were significantly more effective than a third set.

Conducting studies in which participants are observed finding the answers to question in electronic books with only full-text searching or with indexes identified by different algorithms provides an especially promising environment for assessment of techniques for assessing index term quality. It would also be interesting to see whether different techniques for identifying index terms for documents of different length (e.g., journal articles vs. books) or for different domains.

- **Individual and task differences.** A lot of effort has been devoted recently to development of personalized search engines and digital library interfaces (e.g., Khopkhar et al. 2003; Teevan, Dumais and Horvitz 2005; Ma, Pant and Sheng 2007). The results of the experiment described in Wacholder, Liu and Liu (2006) indicate that not-so-good

users (people whose accuracy score was less good on determination of the usefulness of a book for a given topic) did considerably better with print than with PDF; medium of presentation made a difference with a significance level of 0.1. If the distinction between the performance of these two user groups is assumed to reflect individual differences in topic comprehension, these results are evidence that different kinds of users do engage differently in the book selection process.

These results suggest that customizable design of electronic books would also be useful for the book selection process and for access information content. In particular, our results indicate that good readers and not-so-good readers might benefit from differently organized access tools.

In summary, I have described some of my own research on digital books and raised a variety of questions that have arisen about the book selection process, the design and function of content browsing quality, index term quality and customized tools for user needs.

2. Toward a suite of tools for accessing portions of digital books

I suggest that the questions above can be more fruitfully investigated if we envision development of a digital portal that offers a sophisticated suite of inter-related tools specifically designed to help users access book content in non-sequential fashion. Providing such a portal will maximize the affordances of the electronic environment and will be particularly useful in providing customized navigation for non-sequential reading of portions of a book. Many people who want to access digital books do not need to read them sequentially. For example they may be engaged in the book selection process, which requires them to get an overview of the content of the book before they decide whether or not to read it, or they may be looking for discussion of particular issues and themes. It is too soon to envision exactly what this tool will ultimately be like; we are still so tied to the affordances of physical books that it is hard to envision electronic books as anything other than digitized print books. We are also impeded by our familiarity with standard search engines, which were designed primarily to distinguish relatively short documents (at least compared to a book) from other apparently similar documents. The intuition is that by building a suite of such tools and testing its usefulness and usability we create an environment that will facilitate and speed up the process of developing an advanced tool.

Some tools in the suite will help users decide what parts of the book, if any, they want to read; others will make it possible for a user to create a customized overview of the content of a part of the book in the form of a displayed list of index terms; the terms will be geared to user's information needs and knowledge of the field and provide links to related work. Still other tools will display the intellectual context (e.g., works cited; citing works; key authors working the field) of a book or a topic discussed in the book.

Initially, the suite of digital library tools will provide to users a portal to engage in the following operations:

- Browsing, in a non-directed fashion, through a book to get a sense of what it's about.

- Following a particular theme or thread within a book.
- Finding journal articles or other books that cite the book the user is looking at.
- Finding journal articles or other books that address the same issues or questions that the user's book addresses.
- Getting an overview of the content of a part of a book, e.g., a book or a chapter, or a section with a list of high quality displayed terms that can be viewed at different levels of detail and sorted in different ways, e.g., by order within the book, frequency, and possible relevance.
- Finding parts of the book that contain standard or received information.
- Finding books or parts of the book where the author claims to be making points that are novel or original.
- Finding books or parts of books that are primarily for novices or that provide an introduction or an overview to a topic.
- Retracing steps in browsing the book to explore branches of the search path previously not taken.
- Comparing graphs, tables or pictures from different books or different parts of books.

It is already possible to build some of these tools; what is required is investigation of user needs and appropriate adaptation of existing tools to longer documents. Other tools, such as customizable displayed indexes that help the user follow different paths through the book remain a technological challenge.

By setting as our agenda the task of building an integrated set of tools for using electronic books and evaluating the usefulness and usability of these tools, we push ourselves toward creation of an advanced environment that enhances the use of digital books and helps users take maximal advantage of their content.

ACKNOWLEDGEMENTS

This work was funded by NSF grant 0414557, Michael Lesk and Nina Wacholder, PIs.

REFERENCES

- [1] Gorman, M. (2004, December 17). Google and God's Mind: The problem is, information isn't knowledge. *Los Angeles Times*.
- [2] Keen, E. M. (1973). The Aberystwyth index languages test. *Journal of Documentation*, 29(1), 1-35.
- [3] Khopkar, Y., Spink, A., Giles, C. L., Shah, P., & Debnath, S. (2003). Search engine personalization. *First Monday*, 8(7). Retrieved July 21, 2008, from http://www.firstmonday.org/issues/issue8_7/khopkar/index.html.
- [4] Ma, Z., Pant, G., & Sheng, O. R. L. (2007). Interest-based personalized search. *ACM Trans. Inf. Syst.*, 25(1), 5.
- [5] Rice, R. E., McCreddie, M., & Chang, S. L. (2001). *Accessing and Browsing Information and Communication*. Cambridge, MA: MIT Press.
- [6] Salton, G. (1986). Another look at automatic text retrieval. *Communications of the ACM*, 29(7), 648-656.
- [7] Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 449-456). Salvador, Brazil: ACM. doi: 10.1145/1076034.1076111.
- [8] Wacholder, N., & Liu, L. (2006). User preference: A measure of query term quality. *Journal of the American Society for Information Science & Technology*, 57(12), 1566-1590.
- [9] Wacholder, N., & Liu, L. (2008). Assessing term effectiveness in the interactive information access process. *Information Processing & Management*, 44(3), 1022-1031. doi: 10.1016/j.ipm.2007.07.011.
- [10] Wacholder, N., Liu, L., & Liu, Y. (2006). User behavior during the book selection process. *Proceedings of the American Society for Information Science and Technology*, 43(1), 173.