Collection-Level Analysis Tools for Books Online

Catherine C. Marshall
Microsoft Research
1288 Pear Avenue
Mountain View, CA 94043
1.650.693.1308

cathymar@microsoft.com

ABSTRACT

Efforts to develop ebook-based functionality have focused thusfar on the individual artifact—the hardware and software reader—and on promoting user engagement with the page by supporting annotation, clipping, and navigation. In this position paper, I will discuss collection-level analytic functionality that will take advantage of the characteristics of online bodies of literature. This functionality includes (1) wisdom-of-crowds tools; (2) analytic gathering tools; and (3) form- and function-based discovery tools.

Categories and Subject Descriptors

H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *User issues*

General Terms

Design, Documentation, Human Factors

Keywords

Ebooks, annotation, social software, analytic tools, collections

1. INTRODUCTION

In 2001, as the initial surge of ebook euphoria waned, CNI's Director Clifford Lynch warned both sides of the decade-old page versus pixel debate (see [4] for an example of this debate):

> "Given the historic price-performance trajectories for storage, in a few years at least some high-end appliances will house hundreds, if not thousands, of books simultaneously, and certainly laptops with software book readers will house thousands or tens of thousands of books at once. Think of portable personal digital libraries, not portable electronic books, as the future role of these appliances." [7]

This attempt to steer the discussion away from the design of ebooks as isolated artifacts—hardware and software platforms that would serve as the basis for reading an individual book—and toward a larger vision of portable personal digital libraries did not take hold during early ebook research and product development. Instead, a pronounced anxiety about changes in reading fostered a design agenda that focused on making electronic books behave as much like their paper predecessors as possible. An ebook, the story went, should be inviting to read on the screen [2][14]; ebook software should provide the appropriate tools for unselfconscious annotation and for readily clipping passages [16]; ebook hardware should feel like a book in the hand and page-turning should trump the awkwardness of scrolling [6].

And so the ebook narrative went. Most of the interesting ebook capabilities were derived from content analysis algorithms that could support information retrieval, either to find an ebook in the online bookstore or library, or to navigate within an ebook to locate occurrences of a particular word. Cross-document linking—either static or computed—extended the ability to move among books or to look up words. One of the most successful examples of ebook functionality was the Perseus software's ability to show word-by-word translations of ancient Greek texts; indeed many undergraduates found on-demand translation a sufficiently compelling reason to put aside their paper books in favor of on-screen reading [8].

Thus although some attention was devoted to exploring novel capabilities that would take advantage of the computational power (or even the extended storage) offered by ebook platforms, collection-level functionality was not in the limelight. Indeed, initial studies cast doubt on our ability to identify cross-domain functionality that went beyond automating traditional uses of texts [13]. In fact, experimental 'beyond paper' functionality elicited a mix of reactions; successes were mostly domain specific and perception of their utility varied from user to user [10].

But as collections of significant scope and reach become available, it seems worthwhile to once again ask how we can use these collections in ways that we were never able to use the equivalent paper collections, even when they were housed in the rich infrastructure of university libraries or other institutions. As these collections mature, it is time to revisit what it means to go beyond paper, and to ask what a substantial collection of online books can offer in addition to the ability to read on the screens of mobile devices. The portable personal library is compelling as the nexus for traditional research and reference—for reading, for searching, for cross-book linking, for annotation—but perhaps it can also serve as the basis for new kinds of analyses and for new directions in scholarship.

To a certain extent, interest in collection-level analytic tools has been prompted by knowledge discovery and text mining techniques; in the abstract, these techniques promise to support the derivation of new knowledge by bringing together work in different disciplines [18]. Knowledge discovery is arguably a difficult, if potentially exciting, road to take. But there are also lower-road approaches, developing tools that take advantage of characteristics of online books and of their social use.

In this paper, I'll discuss three different directions we can pursue in the name of extending our capacity to work with collections of

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

online books: (1) wisdom-of-crowds tools; (2) analytic tools for gathering; and (3) form- and function-based (as opposed to content-based) discovery tools. Some of these tools are based on current practice, ways scholars and students use texts already; others extend current practice; still others are flights of fancy, but not wholly implausible given directions in textual scholarship.

2. WISDOM OF CROWDS

A wisdom-of-crowds approach takes advantage of the combined actions and judgments of many people, under the assumption that an aggregate opinion of a large group often out-performs a single individual's assessment [17]. Explicit social filtering dates back to the Tapestry system developed in the early nineties [5], and continues to be a well-accepted technique that enables people to select items from large collections based on ratings and recommendations. Many current Web-based services use explicit ratings to recommend particular items: for example, Amazon allows readers to recommend books and other purchases; Netflix uses explicit viewer ratings to suggest movies; and CiteULike[1] applies similar techniques to recommend academic papers. Of course, the authority and motives of one's peers may come into question when one consults individual reviews, but in principle, one over-enthusiastic recommendation is usually moderated by a much larger number of bad or lukewarm reviews and vice-versa.

Other types of recommendation are based on the aggregation of readers' implicit endorsements through actions like sharing or hypertext linking. For example, many periodicals have a 'most emailed' feature to let their readers know what others have found sufficiently interesting to pass around, and many information retrieval rankings use a Google-like notion of popularity based on inbound links to present results in a tractable order [3].

Similar strategies may be applied to select specific segments of text (sentences or phrases, for example) within a longer work. Quotations are an example of explicit recommendation of a passage from a longer work; by analyzing books online, shared quotations may be used to identify significant passages [15].

Implicit within-text selections such as annotations, clippings, or even page views may also be used as the basis for a wisdom-of-crowds approach [9]. These techniques have the distinction of using actions that are unselfconscious and not necessarily directed at an audience (even an audience of oneself in some cases); there is ample evidence that some readers are not aware of how many annotations they have made while they are reading, nor are they always clear on the purpose of their own annotations [11].

Table 1 summarizes item-level and within-text recommendation techniques based on explicit judgments and implicit actions.

Table 1. Explicit and implicit recommendations

	Whole text	*Within-text*
Explicit judgments	Recommending a book; rating a movie	Quoting a passage
Implicit actions	Sharing a newspaper article; linking to a web page	Annotating a selection

Why use these "inside-of-the-book" selections as a surrogate for explicit recommendation of text passages? An individual person's annotations are often little evidence of anything, save a blip in interest, an indication of perceived future utility, an expression of confusion, or some other change in an otherwise smooth engagement with the text. But if multiple people are using the same text for the same thing, there's a good chance that their consensus (in this case, the intersection of the annotations' anchor text) is meaningful. Our study showed that this consensus is significantly more common than would be predicted by a strict probabilistic calculation of anchor overlap [9]. Furthermore, the text that the annotators converge on is often different than the text that authors and publishers designate as important (for example, we can assume that the opening sentence of a section is text that the author thinks is important); in other words, a sentence hidden in the middle of a long chapter may be singled out by many readers through their annotations. If an annotation signals a simple increase in reader attention, then these points of convergence are likely to be more interesting than text that has received less communal attention.

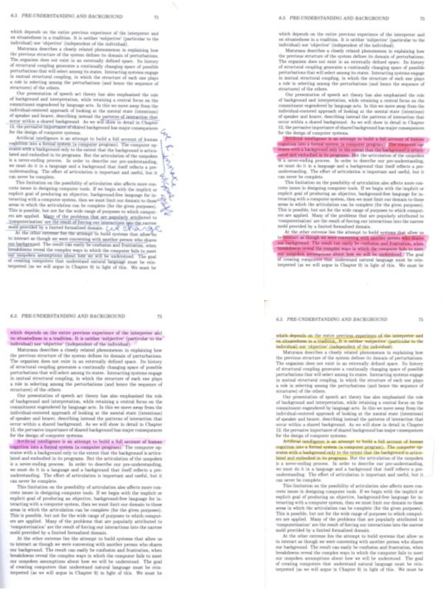
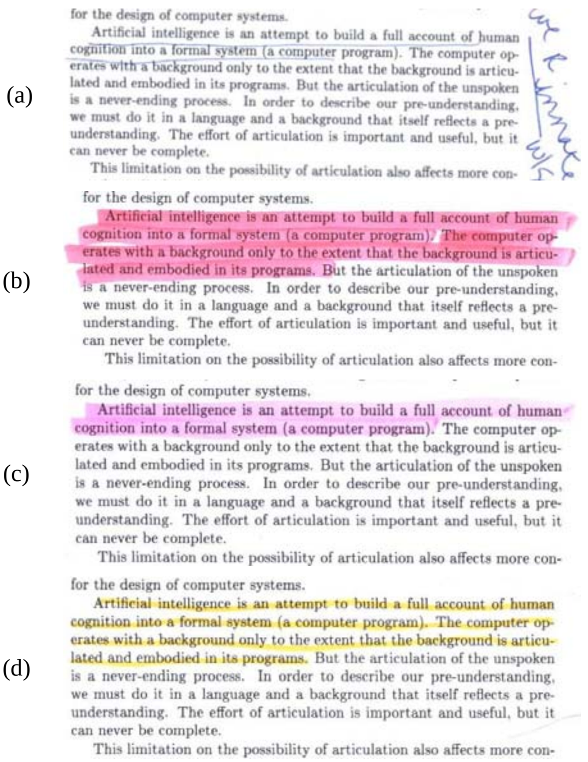


Figure 1. Page 75 from four different readers' copies of *Understanding Computers and Cognition*

Figures 1 and 2 offer a small-scale example of how aggregating annotations enables us to identify interesting sentences within a text. Figure 1 shows thumbnails of page 75 from four copies of *Understanding Computers and Cognition*, a book that was used as a college class text. From this page, it is evident that although their annotations were different, all four readers were interested in the paragraph that is in the middle of the page.

Looking more closely (see Figure 2), it is apparent that all four readers have converged on the first sentence of this paragraph, "Artificial intelligence is an attempt to build a full account of human cognition into a formal system." Two readers (b and d) extend their interest to the second sentence, and only one of the four readers delimits the whole paragraph (using a margin bar) as worth extra attention. Because we are only interested in reader attention, not the specific marks, we only compare the annotations' *anchors* (the extent of text that they specify).

[1] http://www.citeulike.org/



for the design of computer systems.

Artificial intelligence is an attempt to build a full account of human cognition into a formal system (a computer program). The computer operates with a background only to the extent that the background is articulated and embodied in its programs. But the articulation of the unspoken is a never-ending process. In order to describe our pre-understanding, we must do it in a language and a background that itself reflects a pre-understanding. The effort of articulation is important and useful, but it can never be complete.

This limitation on the possibility of articulation also affects more con-

for the design of computer systems.

Artificial intelligence is an attempt to build a full account of human cognition into a formal system (a computer program). The computer operates with a background only to the extent that the background is articulated and embodied in its programs. But the articulation of the unspoken is a never-ending process. In order to describe our pre-understanding, we must do it in a language and a background that itself reflects a pre-understanding. The effort of articulation is important and useful, but it can never be complete.

This limitation on the possibility of articulation also affects more con-

for the design of computer systems.

Artificial intelligence is an attempt to build a full account of human cognition into a formal system (a computer program). The computer operates with a background only to the extent that the background is articulated and embodied in its programs. But the articulation of the unspoken is a never-ending process. In order to describe our pre-understanding, we must do it in a language and a background that itself reflects a pre-understanding. The effort of articulation is important and useful, but it can never be complete.

This limitation on the possibility of articulation also affects more con-

for the design of computer systems.

Artificial intelligence is an attempt to build a full account of human cognition into a formal system (a computer program). The computer operates with a background only to the extent that the background is articulated and embodied in its programs. But the articulation of the unspoken is a never-ending process. In order to describe our pre-understanding, we must do it in a language and a background that itself reflects a pre-understanding. The effort of articulation is important and useful, but it can never be complete.

This limitation on the possibility of articulation also affects more con-

Figure 2. Four-way consensus in four different readers' annotations of *Understanding Computers and Cognition*

Although this example uses the markings of a small number of readers, there is reason to believe that this is an appropriate application of a wisdom-of-crowds strategy to reflect a social consensus of what is interesting in a long text. Surowiecki maintains that the wisdom of crowds is derived from their diversity of opinion, the independence of their choices, decentralization (i.e. local knowledge and context), and the aggregation of individual actions [17]. Aggregating individual anchors rather than trying to extract information from any added content such as marginalia is more likely to identify useful text. Of course, this technique may be trickier for certain genres; readers seldom annotate fiction unless the work is being used as the textbook for a course. However, if we look at the annotations of enough readers over a sufficient duration, it is likely we will find enough annotations to use them in this way.

What can reader consensus do for us? Is there a danger that naïve interpretations will drown out more learned voices, not in the least the authorial voice? Using this approach implies that we take some care about how readers' voices are used. Social consensus can play a role secondary to the text itself: aggregated annotations may, for example, be used as the basis for generating the summaries that are presented with search results; they may be used to automatically emphasize certain text to facilitate skimming, reminding, or to help students study (for example, by comparing their own selections with the selections of others). We can envision different user interface techniques that use progressively widening degrees of interest, presenting the most annotated sentence as the tersest summary of a section, and then expanding that to include the second most annotated sentence and so on. It is crucial to find ways of applying this approach to enhance the text, rather than obscure it.

3. ANALYTIC TOOLS FOR GATHERING

The current way we work with large collections of books online is to search them; existing modes of interaction with the collection are very much oriented to finding a specific book, then reading it, possibly taking notes or annotating it as we read. There are few tools for gathering books or passages based on some kind of analytic criteria. For example, if a college student is taking a Shakespeare course and wants to write about Shakespeare's use of birds and bird imagery in the plays, she is not very much better off than she was in an earlier time. She still lacks a methodical way of gathering the passages that have something to do with birds, and is left to cast about for specific examples, either based on what she remembers from class ("Was there mention of a falcon in *Richard III*? How about a finch in *Midsummer Night's Dream*?") or methodically searching across each play for each bird that crosses her mind (… 'owl' 'osprey' 'ostrich'…).

In fact, given an electronic text, a student is apt to try the latter strategy—searching for instances of the desired word—and be quickly frustrated. The following is a quote from an interview with an English Literature graduate student who was using Microsoft Reader with her course's texts installed on it:

> "I was searching one of the things that [the professor] had suggested ... Things to look for in the critical reading. And just one of the offhand things that she tossed off was, 'do they say anything about universals versus particulars?'… And that caught my interest. … First I searched for 'particular' and he uses the word particular a lot. So, you know, like a 'particular theory', a 'particular couch,' whatever. And not in the sense that I'm looking for…" [11]

She is not getting what she wants 'in the sense that [she's] looking for'. What functionality would support such a critical reading and analysis?

- The ability to limit the scope of a search to a specific subcollection (e.g. just Shakespeare's plays within a larger corpus of English literature);
- The ability to define a thesaurus-like tree of terms to search for (e.g. references to specific types of birds);
- The ability to gather multiple passages together in a space and use tools to work with separate search results in that space (see for example [12]).

Notice that this is partially old-school analysis—we expect considerable human intervention—but there is also a more sophisticated use of technology to scope the search to a portion of the corpus, to perform a controlled expansion of query terms, and to gather the results. The reader is able to start with a relatively complete set of examples of bird references, and is likely to do a more thorough analysis than she would have otherwise.

4. TEXTUAL DISCOVERY TOOLS

Textual discovery tools can exploit aspects of online books that are more straightforwardly available than those used by knowledge discovery tools, for example, the relative prevalence of words. Simple characteristics like word frequency can trace the narrative arc of a novel. For example, Figures 3a and b show a word frequency visualization of two chapters of *Moby Dick*[2]. Figure 3a shows a cloud representing the first chapter of the novel; the predominant words include voyage, sea, water, and time. By the time we get to Chapter 41 (Figure 3b), it is easy to see that the action has come to be dominated by the great white whale. Although these visualizations are playful, textual scholarship may indeed use analyses of such patterns [1]. IBM's Many Eyes information visualization site[3] also offers comparable examples of explorations of texts.

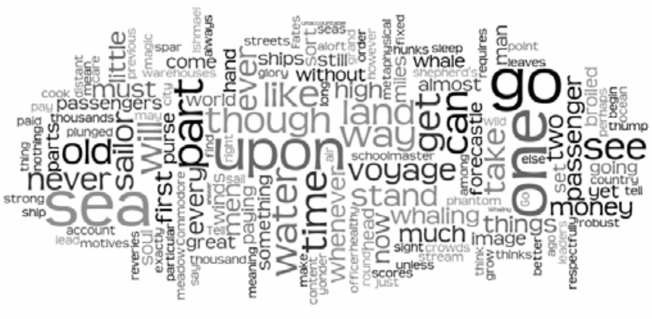


Figure 3a. Melville's *Moby Dick*, Chapter 1, "Loomings" run through Wordle

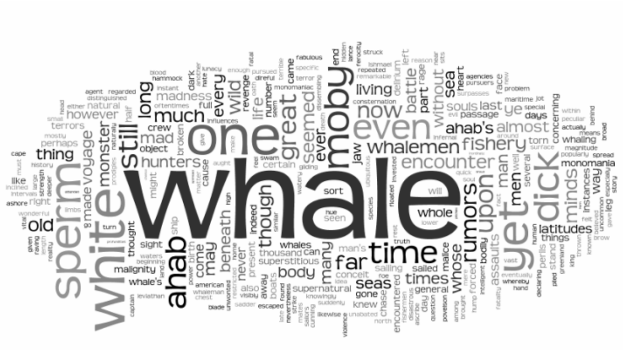


Figure 3b. Melville's *Moby Dick*, Chapter 41, "Moby Dick" run through Wordle

5. CONCLUSION

In addition to forming the basis of personal portable digital libraries, books online provide a unique opportunity to develop collection-level analytic tools. The texts—and records of human interaction with the texts—form a unique dataset that can act as input for further algorithmic analysis and for extracting wisdom-of-crowds recommendations. These results can be gathered and manipulated in promising new ways.

6. REFERENCES

[1] Bernstein, M., Bolter, J.D., Joyce, M., and Mylonas, E. 1991. Architectures for Volatile Hypertext. *Proc. HT'91*. ACM Press, New York, NY, pp. 243–260.

[2] Betrisey, C., Blinn, J. F., Dresevic, B., Hill, B., Hitchcock, G., Keely, B., Mitchell, D.P., Platt, J. C., Whitted, T. 2000. Displaced Filtering for Patterned Displays. *Digest of Society for Information Display Symposium 31*, 1, pp. 296-299.

[3] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A. 1999. Mining the link structure of the World Wide Web. *IEEE Computer 32*, pp. 60-67.

[4] *Feed Magazine*. 1995. Page Versus Pixel. http://web.archive.org/web/19970225062835/http://www.feedmag.com/95.05dialog1.html.

[5] Goldberg, D., Nicols, D. Oki, B., and Terry, D. 1992. Using collaborative filtering to weave an information tapestry, *Communications of the ACM 35*, 12, pp. 61-70.

[6] Liesaputra, V., Witten, I., Bainbridge, D. 2007. Lightweight realistic books. *Proc. JCDL07*, ACM Press, New York, NY, pp. 502-502.

[7] Lynch, C. 2001. The Battle to Define the Future of the Book in the Digital World. *First Monday 6*, 6. http://www.firstmonday.dk/issues/issue6_6/lynch/index.html

[8] Marchionini, G. 2000. Evaluating Digital Libraries: A Longitudinal and Multifaceted View. *Library Trends 49*, 2, pp. 304-333.

[9] Marshall, C.C. 1998. Toward an ecology of hypertext annotation. *Proc. HT'98*, ACM Press, NY, NY, pp. 40-49.

[10] Marshall, C.C., Price, M., Golovchinsky, G., and Schilit, B.N. 2001. Designing e-Books for Legal Research. *Proc. JCDL'01*, ACM Press, New York, NY, pp. 41-48.

[11] Marshall, C.C. and Ruotolo, C. 2002. Reading-in-the-Small: a study of reading on small form factor devices. *Proceedings of JCDL'02*, ACM Press, New York, NY, pp. 56-64.

[12] Marshall C.C. and Shipman, F. 1997. Effects of Hypertext Technology on the Practice of Information Triage. *Proc. HT'97*, ACM Press, New York, NY, pp. 124-133.

[13] McKnight, C. and Dearnley, J. 2003. Electronic Book Use in a Public Library. *Journal of Librarianship and Information Science 35*, 4, pp. 235-242.

[14] Microsoft Reader Press Release. 2008. Downloaded from http://www.microsoft.com/presspass/features/2001/sep01/09-24ebooks.mspx on 20 July 2008.

[15] Schilit, B.N. and Kolak, O. 2008. Exploring a Digital Library through Key Ideas. *Proc. of JCDL08*. ACM, New York, NY.

[16] Schilit, B.N., Golovchinsky, G., and Price, M.N. 1998. Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations. *Proc. CHI98*, ACM Press, New York, NY, pp. 217-226.

[17] Surowiecki, James. 2004. *The Wisdom of Crowds.* Little, Brown.

[18] Swanson, D.R. 1987. Two medical literatures that are logically but not bibliographically connected. *JASIS 38* 4, pp. 228-233.

[2] The visualization was generated by Wordle: http://wordle.net/.

[3] http://services.alphaworks.ibm.com/manyeyes/app