How People Read Books Online: Mining and Visualizing Web Logs for Use Information

Rong Chen[1], Anne Rose[2], and Benjamin B. Bederson[2]

[1] Department of Computer Science and Technique
College of Computer Science, Sichuan University
Chengdu, 610065, China
[2] Human-Computer Interaction Lab
Department of Computer Science
University of Maryland
College Park, MD 20770, USA
chen-rong@cs.scu.edu.cn, {rose,bederson}@cs.umd.edu

Abstract. This paper explores how people read books online using the International Children's Digital Library (ICDL). We analyzed usage of the ICDL in an attempt to understand how people read books from websites. We propose a definition of reading a book (in contrast to others who visit the website), and report a number of observations about the use of the library in question.

Keywords: Web Log Analysis, Information Visualization, Web Usage Mining, ICDL, Reading Online.

1 Introduction

There is now a wide range of online books from sources such as Google Book Search[1], Amazon[2], our own International Children's Digital Library (ICDL)[3], and others. While there is significant effort to understand how people use websites through services such as Google Analytics[4] and various tools to process web logs, these services fall short when trying to understand how people read books online.

The issue is that the existing approaches aggregate data and combine individuals. They support understanding e-commerce activities such as understanding "conversions", and knowing whether certain goals have been reached – such as if a product has been purchased, or whether a book has been downloaded. You can even find out how many pages of some content area have been accessed – so it is possible to discover how many pages of a certain book have been read. But it is impossible using traditional techniques to discover how many individuals have read a book. Or how many pages of a book are typically read by individuals. Or how many books an individual reads. In sum, we want to know how people read books online.

[1] http://books.google.com
[2] http://www.amazon.com
[3] http://www.childrenslibrary.org
[4] http://www.google.com/analytics

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 364–369, 2009.

In this paper, we analyze and visualize web log data. While it would be ideal to actually observe individual reading online, that is not scalable. So, instead we focus on book-centered reading behavior with the actual logs from the ICDL.

This analysis was done on the public usage of the ICDL from one week (20 October 2008 through 25 October 2008), which represents just over 23,000 unique visitors, 26,000 visits, and 336,000 page views.

2 Review of Related Literature

People's online reading behavior has increasingly become an area of empirical and theoretical exploration by researchers from a wide range of disciplines, such as psychology, education, literacy studies, information science and computer science. Different disciplines have diverse ways of probing these questions.

Many researchers use active observation: Some researchers have performed experiments on understanding changes in reading behavior with paper-reading [1][2].

Web page centered research is used by some web usage mining tools [3]. Google analytics gives the average time on page and average reading count of pages in general, but it focuses on each webpage other than each book. So Google analytics can't describe the progress of book reading and how never reports what individuals do.

A number of web log analytics tools exist such as Webalizer, Web Log Expert, Web Log Suite and WUM [4]. However, they are limited in their support for site-wide analysis of the kind we are pursuing to understand how people read books.

3 Visualizing Book Reading Sessions

Because the ICDL is free and allows anonymous usage, relatively few people register with the site. Thus, it is difficult to track an individual's reading progress. But with a bit of effort, we can analyze and track what we call *book reading sessions* (BRS) with reasonable accuracy. We extract BRS as follows.

Step 1: Clean Data: We filtered out records from the Apache web server logs with any error status code as well as records which reference embedded image files.

Step 2: Parse URL: The fields of the web log that we use are IPaddress, Agent, Begintime, Referrer, and URL (which contains fields separated by "&".) For example: /icdl/BookPage?bookid=husblsk_00040002&pnum1=10&pnum2=11&twoPage=true &route=simple_0_0_blue%20sky_English_0&size=0&fullscreen=false&lang=English&ilang=English

Step 3: User Identification & Session Identification: It is a complex process to analyze web logs, but many papers have discussed it [5][6][7]. We follow such identification methods, and define BRS as a time series data set, which includes one "Book Reader" web page (Figure 1) and many "BookPage" web pages (Figure 2).

We can then observe how many pages people read and how much time they spent in each reading session by using visualization software such as LifeLines. LifeLines supports visual exploration of multiple records of categorical temporal data and allows alignment of data on sentinel events, showing intervals of validity [8][9].

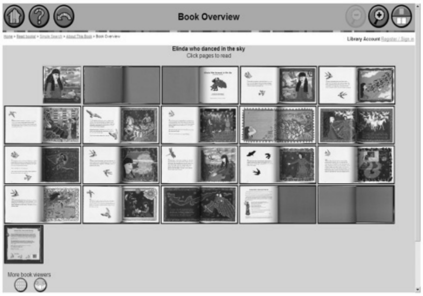
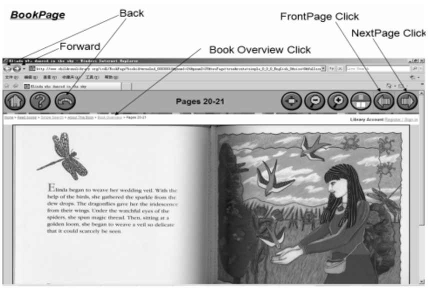

Book Overview
Home • Read books! • Simple Search • About This Book • Book Overview
Library Account Register / Sign in
Elinda who danced in the sky
Click pages to read
More book viewers

Fig. 1. A sample *BookReader* page


BookPage
Back
Forward
Book Overview Click
FrontPage Click
NextPage Click
Elinda who danced in the sky - Windows Internet Explorer
Pages 20-21
Home • Read books! • Simple Search • About This Book • Book Overview • Pages 20-21
Library Account Register / Sign in

Elinda began to weave her wedding veil. With the help of the birds, she gathered the sparkle from the dew drops. The dragonflies gave her the iridescence from their wings. Under the watchful eyes of the spiders, she spun magic thread. Then, sitting at a golden loom, she began to weave a veil so delicate that it could scarcely be seen.

Fig. 2. A sample *BookPage* page

We use LifeLines to visualize BRS (Figures 3, 4). Each *BRS* called a record is vertically stacked on alternating background colors. It is identified by its ID on the left, and its page number ("*Page No*") in this reading session is listed under the session ID in order. Each *BookPage* (called an "event") appears as colored triangle icons on the timeline in the middle of the main display area. The beginning time of the first event (*Page 001*) are aligned vertically.

Seeing this different kind of reading behavior brings us to a key question – what do we mean to "read a book" online? Clearly there are many different styles. Some sessions clearly represent reading and some clearly do not. So, what do we do?

4 Definition of RBRS – Real Book Reading Session

By looking at the BRS data, we now define what we call a Real Book Reading Session (RBRS). It is reasonable to consider a book "read" if every page is looked at for a reasonable time. However, it is difficult to draw a clear boundary between reading and non-reading. For example, if someone reads ¾'s of the pages of a book, while skipping the introductory and ending matter, most people would probably consider that to also be reading the book. What if they skipped two chapters in the middle? Since this is a subjective decision, and our primary purpose was to distinguish people that were doing some reading compared to those that weren't, we decided on a simple and unambiguous definition.

We define a book to be considered read if an individual has looked at more than half of the pages of the book. Therefore, a real book reading session (RBRS) is defined as a book reading session where the percentage of distinct pages read is greater than 50%.

We collected 21,060 sessions, in which 900 books were visited by users (Figure 5). We chose the turning point (50%) in this curve to be the threshold in the definition of RBRS. Based on this definition, 1,197 sessions out of the 21,060 were RBRS, which includes 331 distinct books.

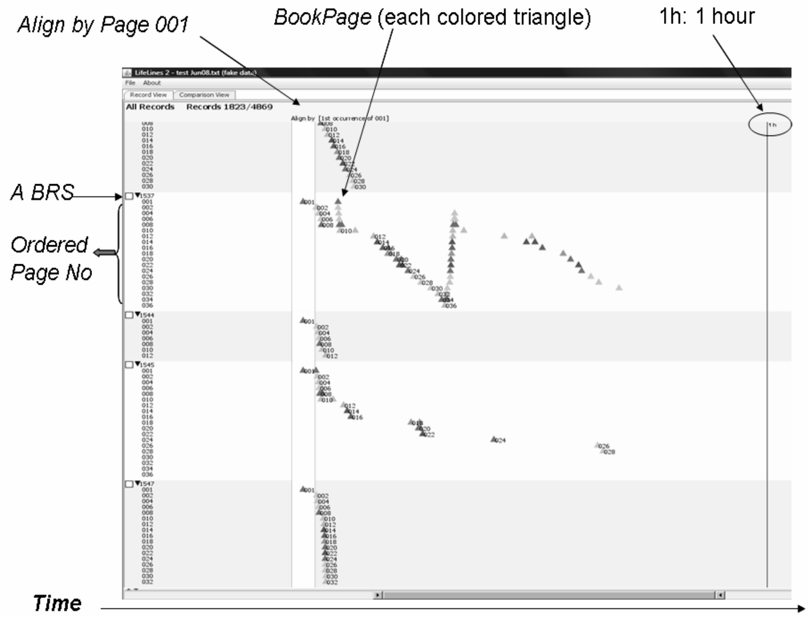

Align by Page 001
BookPage (each colored triangle)
1h: 1 hour
A BRS
Ordered Page No
Time
LifeLines 2 - test Jun08.txt (fake data)
File About
Record View
Comparison View
All Records Records 1823/4869
Align by [1st occurrence of 001]
1h
1537
1544
1545
1547

Fig. 3. This is part of one book's BRSs, includes five book reading sessions .The first BRS on the top spends about 5 minutes on a whole book. The second BRS reads the whole book, then goes back over each page, and reads the entire book a second time more slowly. The third BRS only looks at the first six pages quickly, and then leaves. The fourth BRS looks at the entire book, but there are significant pauses after every few pages. Over one hour is spent on this book. The fifth one looks at every page in the book, but does this so quickly that the entire book is scanned in just one minute.

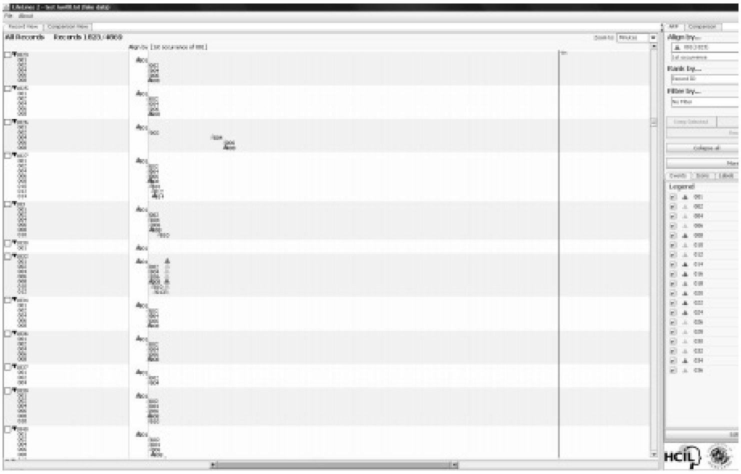


Fig. 4. There are 12 BRSs, each of which includes only a few book pages. They each start at the beginning of the book, and then leave relatively quickly.

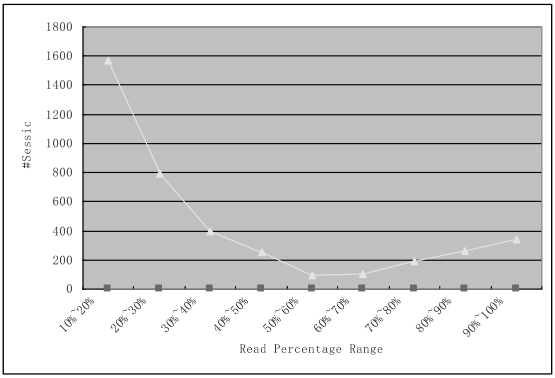


Fig. 5. # *Session* vs. *Read Percentage*. The number of reading sessions (BRS) that represents a user reading the indicated percentage of pages. There is a low value (near 50%) in the figure which motivated us to pick 50% as the number to specify the RBRS cut-off.

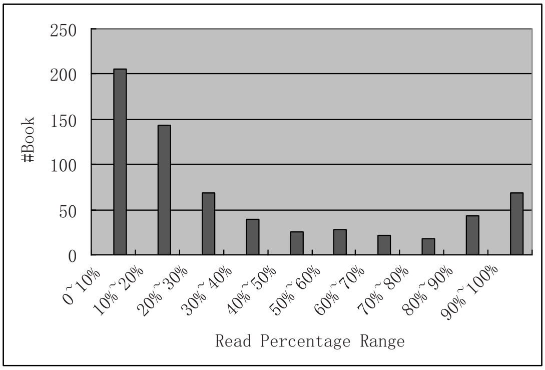


Fig. 6. # *Books* vs. *Read Percentage*. Each bar shows the number of books that had the indicated percentage of pages looked at in a reading session.

5 Web Usage Analysis

We now look at how different books are read. Figure 6 show, as expected, that there are many books which remain mostly unread. And the number of books that have more of them read decreases with the amount of the book that is read – to a point. Surprisingly, the number of books that have 80% or more of the pages read increases. We have no evidence to support an explanation of this. However, we observe that it makes sense that when reading books, it is natural for highly engaged readers to read all of the book – even end paper, etc.

Another thing we looked at was how much time people spent on each page of a book. If people were reading the entire page, then we would expect that the time spent on each page would roughly correlate to the amount of text on that page. To examine this, we picked one book (“Three Little Pigs”) that had a varying amount of text on each page that was read relatively often. The results are somewhat surprising in that while the time spent on each page is clearly lower when there are very few words on a page, it clearly does not increase directly with the amount of text on each page. Again, we don't have any data that explains this, but some possible explanations are that many readers may be simply looking at pictures. Another possibility is that our hypothesis is incorrect, and people naturally spend a roughly constant amount of time per page – perhaps looking at pictures more, or reading more slowly when there is less text.

In sum, this short paper just scratches the surface in looking at how people actually read books in the ICDL. The data indicates some expected, but some surprising results – which clearly indicate the need to study this in more detail.

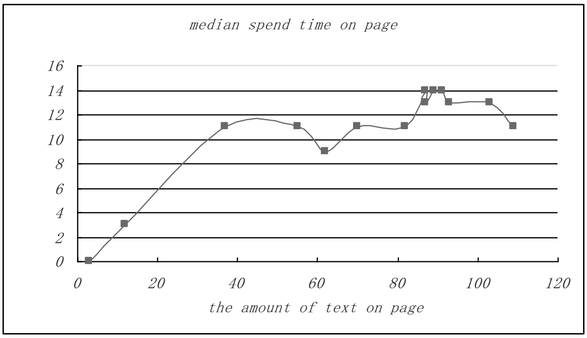


Fig. 7. This shows the *median spent time* (in seconds) correlated with *the amount of text on page* (in words), for the book "Three Little Pigs"

References

1. Liu, Z.: Reading behavior in the digital environment Changes in reading behavior over the past ten years. Journal of Documentation 61, 700–712 (2005)
2. O'Hara, K., Sellen, A.: A comparison of reading paper and on-line documents. In: Proceedings of Human Factors in Computing Systems(CHI 1997), pp. 335–342. ACM Press, New York (1997)
3. Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In: ACM SIGKDD Explorations Newsletter. ACM Press, New York (2000)
4. http://hypknowsys.sourceforge.net/wiki/The_Web_Utilizat
5. Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. J. Knowledge and Information Systems 1(1), 5–32 (1999)
6. Wu, K.-L., Yu, P.S., Ballman, A.: Speed tracer: a web usage mining and analysis tool. IBM Systems Journal 37(1), 89 (1998)
7. Yang, Z.L., Wang, Y.T., Kitsuregawa, M.: An Effective System for Mining Web Log. In: Zhou, X., Li, J., Shen, H.T., Kitsuregawa, M., Zhang, Y. (eds.) APWeb 2006. LNCS, vol. 3841, pp. 40–52. Springer, Heidelberg (2006)
8. Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B.: Lifelines: visualizing personal histories. In: Proc. CHI (1996)
9. Wang, T.D., Murphy, S., Plaisant, C.: Aligning Temporal Data by Sentinel Events: Discovering Patterns in Electronic Health Records. In: Proc.CHI 2008. ACM Press, New York (2008)