Ten Months of Digital Reading

An Exploratory Log Study

Pavel Braslavski[1], Vivien Petras[2], Valery Likhosherstov[1], and Maria Gäde[2]

[1] *Ural Federal University | pbras@yandex.ru ; v.lihosherstov@gmail.com*
[2] *Humboldt-Universität zu Berlin | maria.gaede@ibi.hu-berlin.de ; vivien.petras@ibi.hu-berlin.de*

This is an author's accepted manuscript version of a conference paper published in the International Conference on Theory and Practice of Digital Libraries TPDL 2016: Research and Advanced Technology for Digital Libraries within the Springer Lecture Notes in Computer Science book series (LNCS, volume 9819).

The final publisher's version is available online at:
https://doi.org/10.1007/978-3-319-43997-6_32

Ten Months of Digital Reading: An Exploratory Log Study

Pavel Braslavski[1], Vivien Petras[2], Valery Likhosherstov[1], and Maria Gäde[2]

[1] Ural Federal University pbras@yandex.ru, v.lihosherstov@gmail.com
[2] Humboldt-Universität zu Berlin {maria.gaede, vivien.petras}@ibi.hu-berlin.de

Abstract. We address digital reading practices in Russia analyzing 10 months of logging data from a commercial ebook mobile app. We describe the data and focus on three aspects: reading schedule, reading speed, and book abandonment. The exploratory study proves a high potential of the data and proposed approach.

Keywords: digital reading; reading behavior; user modeling; log analysis

1 Introduction

The volumes of digital reading have been steadily growing several years until recently. According to Pew Research, the share of Americans who read at least one ebook yearly grew from 17% in 2011 to 28% in 2014; there were 4% of "ebook only" readers in 2014.[3] In Russia, where our data comes from, the figures for spring 2015 are quite similar: 25% of adults have read fiction ebooks at least once in the last 12 months; 8% read ebooks at least once a week.[4]

Distinct ebooks features can potentially influence everyday reading patterns and behaviors. For example, ebooks provide search, multimedia and hypertext functionalities that cannot be implemented in print. Another characteristic is content accessibility: a small mobile device can give the reader instant access to hundreds of thousands of book titles. Especially with streaming content delivery model (i.e. the user pays a flat rate and gets access to the whole collection) as studied in this paper, we assume that ebook readers follow a "try-and-drop" scenario more often.

This paper reports on an initial log study of ca. 3 million reading sessions of about 8,000 users during 10 months of 2015 from a mobile application in Russia. This work-in-progress demonstrates the unprecedented opportunities for low-level analysis of reading at scale that cannot be conducted based on surveys, controlled user studies, and book-level consumption data. In this paper, we describe the dataset and demonstrate its potential focusing on three aspects: 1) reading schedule, 2) reading speed, and 3) book abandonment. We are not

[3] http://pewrsr.ch/1LZOwBb
[4] http://www.levada.ru/2015/05/19/rossiyane-o-chtenii/(in Russian)

aware of other large-scale reading log studies, except for solitary self-reported data from ebook services. Since reading is a central cultural practice, reading log analysis can be beneficial for many domains, e.g. schooling, second-language learning, creative writing, book publishing and recommendation, etc.

2 Related Work

Digital reading has been studied from HCI, educational and psychological perspectives. Previous research dealing with reading online or ebooks has focused on differences between screen and paper reading behavior as well as on contextual motivations, preferences and technological challenges in work-related or casual leisure situations [1,2,4,7]. Usage data of ebooks was studied focusing mainly on ebook selections or retrieval issues [3,5,12].

Due to the complexity of the reading process, studies investigating reading and in-book navigation patterns or reading strategies are rather underrepresented [8]. Some researchers report overlapping reading patterns that vary from linear, browsing to berry picking ebook sessions [9,13]. Comparing the contradictory results, study parameters and contextual issues seem to play an important role dealing with reading behavior [2].

The majority of research dealing with ebooks has focused on academic settings in English speaking countries [10] and used qualitative data such as interviews, diaries or observations focusing on individual differences and preferences reading online. While the usage logs of digital library ebooks have been investigated [11,6], non-academic genres are mainly represented by sales rates[5].

3 Bookmate Data

Bookmate[6] is a popular Russian digital reading service. Upon installing an application, users get instant access to the free collection. Standard paid subscription grants a user access to the entire Russian book collection, excluding new arrivals, bestsellers, and business books. Premium subscription provides unlimited access to the entire Bookmate collection. Bookmate logs used in the study correspond to almost 10 months – from January 1st to October 22nd, 2015. The data includes information about the users, books, and readings sessions.

Users. Title preferences of paying and non-paying subscribers are remarkably different – the latter seem to focus on classical novels, mostly by Russian authors. We speculate that these might be required reading material for high-school literature classes, indicating different reading behaviors than the general public. To reduce the variation in the sample, we focus our analysis on the behavior of the paying Bookmate users, who spent more than five hours in the app and read at least 10 different books during the study period. We refer to this group

[5] `http://www.theguardian.com/media/2012/feb/05/ebook-sales-downmarket-genre`
[6] `https://www.bookmate.com`

of 8,337 users as *CORE_USERS*. 6,897 (83%) of them indicated their gender; there is an almost equal number of female (3,445) and male (3,452) readers in the population. Out of 2,804 (34%) *CORE_USERS* who indicated their year of birth, the majority were born in the 1980s (51%) and 1990s (28%).

Reading sessions. Approximately 172 million interactions were recorded for the *CORE_USERS*. Interactions contain user and book IDs, time stamps, and the character ranges that the user read or just browsed through in a certain book. Single interactions were aggregated into reading sessions comprising all subsequent interactions for one user with less then a 30 minute pause between them.[7] This resulted in 3.1 million reading sessions. In addition, we isolated 'fast-forward' (faster than 300 words/min) and backward browsing sessions as navigational and did not consider them in statistics dealing with reading volumes. The median value of 305 sessions corresponds roughly to a daily usage pattern.

Books. The Bookmate collection contained 523,689 ebooks by the end of 2015. Roughly half of all books (243,264) are categorized according to 20 genre labels, such as *Love & Romance*, or *Politics & Society*. The book collection read by the *CORE_USERS* consists of 72,823 items; 10,316 of them are read by at least ten readers. It has to be noted that Bookmate's books do not always correspond one-to-one to printed editions: ebooks range from short verses (several hundred characters) to e-versions of multi-volume collections.

4 Digital Reading Behaviors

4.1 Reading Schedule

Reading logs allow us to uncover reading schedule at different scales: hours, days, and weeks. Fig. 1 shows average weekly reading volumes over the entire period of observations. Although these volumes can be affected by instability of the user base and promotional campaigns, we can see a higher activity in the period of New Year holidays (January 1–11) and vacation season (July and August). There is also a noticeable increase in reading activity during spring holidays (first decade of May). This indicates more leisure-time reading pattern. Reading activity during the week shows a curious pattern: it increases from Monday (100%) to Wednesday (102.9%) and then drops, reaching the lowest point on Saturday (93,2%). We can speculate that on Saturdays some Bookmate users prefer other leisure activities than reading. Fig. 2 shows that most reading activities occur in the evening and night, which again corresponds to leisure reading pattern. Minor differences in fiction/non-fiction reading during the day are rather expected: non-fiction's relative figures surpass fiction's in the

[7] We adopted 30-minute threshold widely used in search query log studies.

morning (10am–12pm), while fiction wins back afternoon and in the late night.

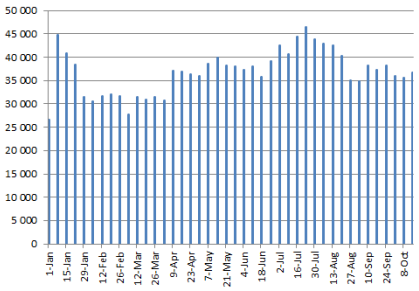

Fig. 1. Average user weekly reading volume.

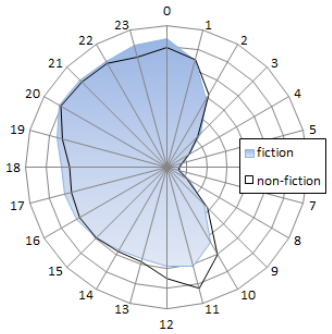

Fig. 2. Relative volumes read by hours of the day for fiction vs. non-fiction.

4.2 Reading Speed

Log-based analysis of reading speed can be seen as a large-scale reading proficiency test (when focusing on the users) and readability study (when looking at the books). Users' reading speeds have a bell-shaped distribution with mean around 150 words/min, which is the upper bound of the recommended reading speed for elementary school graduates. Fig. 3 ranks genres by reading speed and thus reflects their 'difficulty', with recipes being most 'readable' and poetry – least easy to read. It is interesting to note that kids books that are expected to have an easy writing style appear in the 'difficult' subspectrum. This suggests that these books are either read by kids themselves or by parents aloud.

4.3 Book Completion and Abandonment

Reading logs provide us with an exceptional opportunity to determine book completion rates. It is also interesting to validate our hypothesis about higher abandonment rates for ebooks under streaming subscription models, although direct comparisons with printed books are hardly feasible.

We consider the book abandoned if a user does not recur to it within one month after the last reading (thus, a final reading must take place one month before the end of the period presented in the data). In addition, we require that the user started to read the book (i.e., there is a session corresponding to the first 10% of the book) within our observation period. As a result of these limitations, we have 534,200 unique user–book pairs, and only 190,879 (35,7%) have completion rate above 90%. This score is lower than what was reported by the Kobo service that sells individual ebooks.[8] Another figure on book completion can be found in the Goodreads survey: 38.1% of users reported

[8] `http://nyti.ms/1QFpcWz`

that they always finish the book, when started.[9] Bookmate users show much lower persistence: only 111 (1.3%) of them read at least 90% of their books until 90% of their length and beyond.

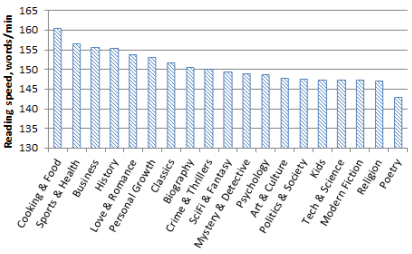

Fig. 3. Reading speed averaged over genres.

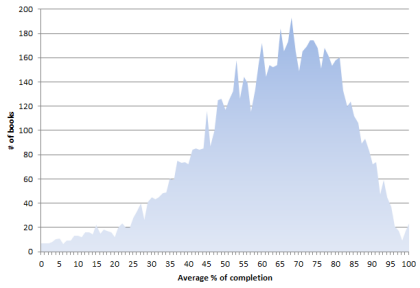

Fig. 4. Numbers of books read until the specified percent of their length.

In addition, we calculated the average point of abandonment for a subset of popular books (8,274 books with 10+ readers that meet additional requirements stated above), distributed as shown in the Fig. 4. About half of the books is abandoned upon reading 64% or less of their length; only about 5% of the books are read beyond 90% of their length. For example, averaged completion rates for *Fifty Shades of Grey* by E. L. James, *Bridget Jones's Diary* by Helen Fielding, and *Martian* by Andy Weir are 69%, 71%, and 78% of book length, respectively. It is interesting to note that the book length does not correlate with its completion rate.

5 Conclusion and Future Work

In the study, we described the Bookmate application log that corresponds to 10 months of reading of about 8,000 users and provided some initial statistics reflecting three aspects of reading behavior: 1) reading schedule, 2) reading speed, and 3) book abandonment. The main results of the current exploratory study are following:

- Bookmate users' reading corresponds to the leisure-type activity; reading is increasing during holidays and vacations; a moderate decrease on Saturdays can be associated with alternative leisure activities practiced on weekends; reading sessions are shifted towards evening and night hours;
- Ebook reading speeds correspond seemingly to those of printed books; log-based reading speed analysis enables large-scale readability and reading proficiency user studies;

[9] `http://bit.ly/1RAM32Y`

– Flat-rate subscription model seems to promote "try-and-drop" pattern and lower book completion rates. Indirect comparisons suggest that these behaviors differ both from those of printed books readers and readers purchasing individual ebooks.

In our future work we will elaborate on the described aspects and address other characteristics of reading behavior, including title preferences and reading style (one-by-one vs. interleaved book reading), as well as investigate book genre and reader's gender facets. Moreover, we will attempt to map low-level navigational and reading interactions to book content, which opens new opportunities to reading analysis. The analyses can be beneficial for different domains and applications: digital libraries, creative writing and book publishing, as well as book recommendation.

6 Acknowledgments

We thank Bookmate and Samer Fatayri in person for preparing the dataset and granting access. Our study was performed using the computational cluster of the Ural Federal University.

References

1. A. Adler et al. A diary study of work-related reading: Design implications for digital reading devices. In *CHI'98*, pages 241–248.
2. G. Buchanan, D. McKay, and J. Levitt. Where my books go: Choice and place in digital reading. In *JCDL'15*, pages 17–26.
3. A. Hinze et al. Book selection behavior in the physical library: implications for ebook collections. In *JCDL'12*, pages 305–314.
4. A. Hupfeld, A. Sellen, K. O'Hara, and T. Rodden. Leisure-based reading and the place of e-books in everyday life. In *INTERACT'2013*, pages 1–18.
5. J. Y. Kim, H. Feild, and M. Cartright. Understanding book search behavior on the web. In *CIKM'2012*, pages 744–753.
6. H. Littlewood et al. A log analysis study of 10 years of ebook consumption in academic library collections. In *ICADL'2014*, pages 171–181.
7. Z. Liu. Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *J. Doc.*, 61(6):700–712, 2005.
8. C. C. Marshall. Reading and writing the electronic book. *Synthesis lectures on information concepts, retrieval, and services*, 1(1):1–185, 2009.
9. D. McKay. A jump to the left (and then a step to the right): Reading practices within academic ebooks. In *OzCHI'11*, pages 202–210.
10. J. Staiger. How e-books are used: A literature review of the e-book studies conducted from 2006 to 2011. *Ref. User Serv. Q.*, 51(4):355–365, 2012.
11. J. C. Tucker. Ebook collection analysis: subject and publisher trends. *Collection Building*, 31(2):40–47, 2012.
12. C. Willis and M. Efron. Finding information in books: Characteristics of full-text searches in a collection of 10 million books. *Proc. of ASIST*, 50(1):1–10, 2013.
13. T. Zhang and X. Niu. Final report for 2015 er&l + ebsco library fellowship research project. In *Libraries Reports, Paper 4*, 2015.