Large-Scale Log Analysis of Digital Reading

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

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

This is an author's accepted manuscript version of a conference paper published in *Creating Knowledge, Enhancing Lives through Information & Technology. Proceedings of the 79th Association for Information Science and Technology (ASIS&T) Annual Meeting (Vol. 53, pp. 1–10), October 14-18, 2016, Copenhagen, Denmark.*

The final publisher's version is available online at:
https://doi.org/10.1002/pra2.2016.14505301044

Large-Scale Log Analysis of Digital Reading

Pavel Braslavski
Ural Federal University
Yekaterinburg, Russia
pbras@yandex.ru

Vivien Petras
Humboldt-Universität zu Berlin
Berlin, Germany
vivien.petras@ibi.hu-berlin.de

Valery Likhosherstov
Ural Federal University
Yekaterinburg, Russia
v.lihosherstov@gmail.com

Maria Gäde
Humboldt-Universität zu Berlin
Berlin, Germany
maria.gae de@ibi.hu-berlin.de

ABSTRACT
In this paper, we address daily reading practices of the general public in Russia analyzing 10 months of log data from the commercial ebook site Bookmate. We study different reading characteristics with ebooks, i.e. the reading volume and preferences, reading schedule, reading speed and reading style (including parallel reading patterns and book abandonment rates), with respect to reader gender, book length and genre of the book. We find that book genres impact certain reading behaviors, while gender differences or book length seem to play less of a role in ebook reading. Parallel book reading and book abandonment occur very frequently, possibly pointing towards changing reading behaviors in the ebook environment. The obtained insights demonstrate the high potential of log analysis for book reading studies.

Keywords
Ebooks, digital reading, reading behavior, user modeling.

INTRODUCTION
Several years ago, Umberto Eco argued that the computer cannot compete with a printed book, because the latter can be read while lying in a bathtub (Eco & Carrière, 2009).

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

ACM ISBN 987-1-4503-2138-9
DOI: 10.1145/1235

Digital reading has been steadily growing for several years now. 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[1].

In Russia, where our data comes from, the figures for spring 2015 were 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[2].

Ebooks are different from their printed counterparts in several aspects. An ebook is much more `fluid' compared to traditional paper books: there is no pagination, fixed font size or type; ebook adapts itself to different screen sizes and resolutions. Ebooks can contain various multimedia content and enable alternative access strategies, from searching within the full-text to following built-in recommendations to other books. Another distinctive ebook feature -- a small mobile device can give the reader instant access to hundreds of thousands of book titles -- is incomparable in its affordances even to visiting a library or a large bricks-and-mortar bookstore. These ebook features can potentially lead to changes in book consumption and reading behavior that become more evident with the advances in ebook technologies and the subsequent new interactions that they make possible.

Ebooks have another intriguing characteristic: like any other content in electronic form (music, movies, web pages, search queries, etc.), they allow for tracking readers' behaviors at unprecedented scale and granularity level and as such open new opportunities for reading analysis.

This paper reports the results of a log study of 10 months of reading data from a commercial ebook subscription site in Russia. Over 3 million reading sessions of ca. 8,000 readers are studied. The study explores the potential of reading logs

[1] http://pewrsr.ch/1LZOwBb
[2] http://www.levada.ru/2015/05/19/rossiyane-o-chtenii/

and provides insight into reading behaviors of a broad range of readers. As it is notoriously difficult to study data from commercial sites, we are not aware of other large-scale datasets or studies based on them, except for self-reported data from the book selling/subscription services themselves.

Our overall research objective is to identify the aspects that affect different reading practices and the indicators that determine these behaviors. The research question is:

- "What are the differences in digital reading behaviors on a commercial ebook website compared to print reading or digital reading behaviors in academic situations?"

We hope to gain more large-scale evidence about reading behavior such as reading preferences, schedule, and volumes. Especially in the case of a streaming content delivery model (i.e. the users pay a flat rate and get access to the whole collection) as is studied in this paper, we assume that the ebook readers follow a "try-and-drop" scenario more often, because so many books are available instantaneously and simultaneously. Based on log data, this work demonstrates the opportunities for 'low-level' analysis of reading at scale that cannot be conducted based on surveys, controlled user studies, and book-level consumption data. We gain access to much finer-grained information that is very hard to obtain *en masse* for printed books: reading speed, book completion and abandonment, etc.

Because of the differences in affordances between digital and print books and in user contexts between the subscription-based ebook site used in this study and the laboratory or library settings prevalent in other studies, it might be difficult to draw a direct comparison of reading practices. Nevertheless, the comparisons will show first trends where digital reading research can continue.

We hypothesize four factors that could have an impact on different reading behavior and interactions:

- the reader, e.g. gender, age, education level, reading level, geographic location;
- the format and other formal features of a book, e.g. length of a book, publication date, language;
- the content of a book, e.g. genre, the amount of multimedia content included, linear storytelling or encyclopedic, linked components;
- the reading environment, e.g. type of e-reader, place of reading (i.e. on the move on a mobile device), possible distractions (i.e. other applications, people), etc.

In this paper, we concentrate on the first three aspects listed above: comparing reading behaviors by the gender of readers, by length of the book and by genre. Our interpretations of reading behaviors are based on indicators calculated from the reading sessions provided. Because this study is based on a logfile of user interactions, the user motivation for certain behaviors remains inconclusive. However, certain behavioral patterns can be determined, which will be followed up by more detailed qualitative studies in future work.

In this paper, we will analyze indicators for different digital reading behavior characteristics, which will be detailed later in the paper:

- reading volume & preferences,
- reading schedule,
- reading speed, and
- reading style, which includes parallel reading and book completion & abandonment.

Reading is largely determined by cultural, economic, and educational traditions. However, some preferences seem to cross national and cultural boundaries. In our study, we see that the readers favor books that are Russian translations of international bestsellers, ebooks adoption rates is Russia is similar to those in the USA, and mobile devices are the same. This suggests that the results of the study are not restricted to Russia, but could be valid for other markets as well.

The paper is an extended version of a preliminary analysis (Braslavski et al., 2016).

The paper is organized as follows. Section 2 describes related work on reading behaviors. Section 3 describes the dataset, its sampling and structure. Section 4 presents the results according to the analyzed characteristics and discusses the results in comparison to other studies. Section 5 concludes the paper and discusses future work.

RELATED WORK

Reading and in particular 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 (Adler et al., 1998; Buchanan et al., 2015; Hupfeld et al., 2013). While an early study found no significant differences in reading speed and comprehension between digital and paper versions of books (Oborne and Holton, 1988), newer research reports differences in reading fluency (faster reading speed), involvement and understanding (less reading comprehension) of digital texts compared to print (Akbar et al., 2015; Murphy et al., 2003; Wagner et al., 2012). Lui (2005) found that people spend more time reading and browsing texts online than they did before reading paper based only. Also, ebook presentations often seem to pose challenges with respect to navigational issues (Malama et al., 2004). The "physiology" of reading appears to be the same no matter the medium: Zambarbieri and Carnigliaa show that reading printed books vs. ebooks does not vary significantly in terms of eye movements (Zambarbieri and Carniglia, 2012).

Current research monitors ebook reading behavior in controlled experiments with particular audiences, for example school students (Hwang and Yueh-Min, 2014; Simpson et al., 2013), or in particular settings and domains, for example libraries (Littlewood et al., 2014).

Usage data of ebooks was studied focusing mainly on ebook selections (i.e. what factors contribute to a reader choosing a particular book) or retrieval issues (i.e. what components of the book are helpful for search) (Hinze et al., 2012; Kim et al., 2012; Willis and Efron, 2013). Within the digital library domain, ebook borrowing data reveals a variety of behavior patterns ranging from simple printing services to more extensive reading and borrowing sessions (McKay and Buchanan, 2016).

Gender and age differences in reader were identified with respect to reading skills, reading navigation, information seeking and technology issues (Huang et al., 2013; Liu and Huang, 2008). Survey responses indicate that men feel more comfortable with ebooks while women state to be the more active readers (Rowlands et al., 2007). Comparing Chinese students, a study reveals that women tend to read linear while men are much more selective, spending more time on browsing and scanning (Liu and Huang, 2008).

Due to the complexity of the reading process, studies investigating reading and in-book navigation patterns or reading strategies are rather underrepresented (Marshall, 2009). Some researchers report overlapping reading patterns that vary from linear (from the beginning to the end), browsing (page hopping) and berry picking (selective searching for information or references) (McKay, 2011; Zhang and Niu, 2015).

While these studies are encouraging, many report results that are difficult to generalize or contrast with previous results. Study parameters and contextual issues seem to play an important role when analyzing reading behavior (Buchanan et al., 2015).

The majority of research dealing with ebooks has focused on academic settings in English speaking countries (Staiger, 2012) 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 (Tucker, 2012; Littlewood et al., 2014), non-academic genres are mainly represented by sales rates[3].

BOOKMATE DATA

Bookmate[4] is a popular Russian digital reading service. The service is similar to the US-based Oyster[5] and German Skoobe[6] ebook sites. Subscribers pay a monthly flat rate and are granted access to the entire book collection in contrast to ebooks stores like Amazon[7] or the Russian Litres[8], where users buy or rent individual books at a time. Bookmate users get access to ebooks through mobile apps and can download content to their devices for offline reading.

Upon installing an application, users get instant access to the free collection (about 7,450 titles by the end of 2015). Some titles are only temporarily free-of-charge for promotional purposes. 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.

Bookmate Users

Title preferences of paying and non-paying subscribers are remarkably different (see Table 1). 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 focused our analysis on the paying users who spent more than five hours in the app and read at least 10 books during the logging period in order to represent a group that shows significant and continued usage of the site. During the 10 months of data collection, there were 15,808 unique readers fulfilling the frequency criterion; 8,337 of them are paying. We refer to the latter group as *CORE_USERS*.

Many *CORE_USERS* (6,897, 83%) indicated their gender; there is an almost equal number readers marked their accounts as 'female' and as 'male' – 3,445 and 3,452, respectively. Out of 2,804 (34%) *CORE_USERS* who indicated their year of birth, the majority were born in the 1980s (51%) and 1990s (28%)[9].

A subset of reading sessions contains geographic locations based on IP addresses. The available data shows that 79.7% of the *CORE_USERS* are from the two largest Russian cities – Moscow and Saint Petersburg.

The distribution of mobile platforms shows a clear preference of users for Apple products (70.1%), followed by Android-based devices (25.7%) and Windows-based

[3] http://www.theguardian.com/media/2012/feb/05/ebook-sales-downmarket-genre

[4] https://www.bookmate.com

[5] https://www.oysterbooks.com

[6] https://www.skoobe.de/

[7] http://www.amazon.com

[8] http://www.litres.ru/

[9] According to the 2010 Russian census, these groups correspond to 16% and 14% of the entire country population, respectively; see http://perepis-2010.ru/.

Non-paying users	Paying users
Aleksandr Kuprin *The Garnet Bracelet*	Donna Tartt *The Goldfinch*
Andy Weir *Martian*	Boris Akunin *Planet Water*
Mikhail Bulgakov *The Master and Margarita*	Max Bazerman *The Power of Noticing*
Ray Bradbury *Fahrenheit 451*	Andy Weir *Martian*
Anton Chekhov *The Cherry Orchard*	E. L. James *Fifty Shades of Grey*
Maxim Gorky *The Lower Depths*	Tiina Orasmae-Meder *Beauty Myths*
Anton Chekhov *Ward No. 6*	Eleanor Catton *The Luminaries*
Anton Chekhov *About Love*	Sergei Lukyanenko *Sixth Watch*
Leo Tolstoy *War and Peace*	Jojo Moyes *Foreign Fruit*
Fyodor Dostoyevsky *Crime and Punishment*	Ayn Rand *Atlas Shrugged*

Table 1. Top 10 books among non-paying vs. paying users.

devices (3.4%). Interestingly, this distribution differs from mobile web browsing usage in Russia by the end of 2015, where Android-based devices cover 71.6% of the market share, followed by iOS (26.2%) and Windows Phones (2.2%)[10]. The most popular devices used by the *CORE_USERS* are iPhones or iPads. Based on other statistics that show that iPhone users generally have a higher income than Android users[11], we postulate that Bookmate subscribers might be more affluent than the Russian populace in general (based on the device usage).

The size of the user population in the study is much larger than any reported in previous digital reading studies. However, this group does not represent the general Russian citizen as users are mainly residents of the two largest Russian cities, younger than 40 years and more affluent than average. At the same time, this group of early adopters provides insights into the behavior of general population in the near future, with the further evolvement of digital reading.

The Bookmate Book Collection

The Bookmate collection contained 523,689 ebooks by the end of 2015. Almost three quarters (385,265) of them had not been read by any of the subscribers during the time of observation. The majority of the books are in Russian (245,262) and English (210,759). The collection size is comparable to other Russian digital book collections. For instance, the catalog of Litres, the leading seller of ebooks in Russia, contained ca. 160,000 items by the end of 2015, whereas OZON[12], the leading online print book seller, offered more than 560,000 unique books (including print-on-demand books) by the same time. In this paper, we concentrate on a core group of users, which reduces the number of read books in our sample to 72,823. These are referred to as *CORE_BOOKS* hereafter.

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[13].

Bookmate offers a wide range of ebooks, both in terms of genre and length. The distribution of *CORE_BOOK* lengths is presented in Figure 1. Shorter books of less than 100,000 characters (length calculated in characters, because pages are variable in ebooks) occur most often, other sizes are evenly distributed.

Roughly half of all books (243,264) are categorized according to 20 genre labels, such as *Love & Romance*, or *Politics & Society*. A book may have several genre labels, in some cases belonging simultaneously to fiction and non-fiction genres. Most books in the studied sample are fiction books (see Figure 2).

Reading Sessions

Approximately 172 million interactions were recorded for the *CORE_USERS*. Every recorded interaction in the Bookmate reading log contains the user and book IDs, the time stamp, and the character ranges that the user read or just browsed through in a certain book. If the user takes no action within two seconds, an interaction record is generated and sent to the server. When the reader goes offline, the reading actions are stored on the device and are sent to the server, once the reader goes online.

Single interactions were aggregated into reading sessions comprising all subsequent interactions for one user with less than a 30 minute pause between them[14]. This resulted in 3.1 million reading sessions. In addition, we isolated `fast-forward' (faster than 300 words/min) and backward

[10] See live statistics: http://www.liveinternet.ru/stat/ru/oses.html?slice=mobile
[11] http://bit.ly/1ZIqoj8
[12] http://www.ozon.ru/

[13] The longest item in the collection is Sergey Solovyov's 29-volume *History of Russia from the Earliest Times* (23M+ characters).
[14] We adopted a 30-minute threshold widely used in search query log studies.

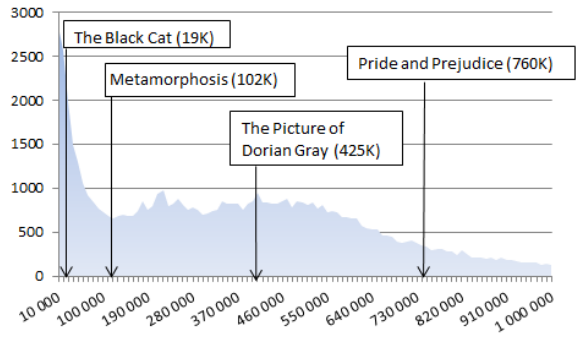

Figure 1. Distribution of book lengths in the *CORE_BOOKS*. The figure only shows books with less than 1M characters; longer books comprise 6.7% of *CORE_BOOKS*.

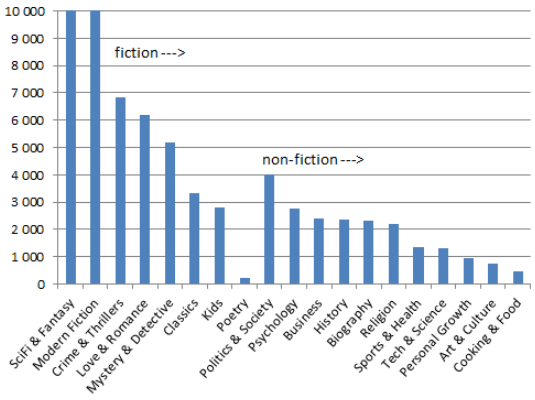

Figure 2. Distribution of genres in *CORE_BOOKS*.

browsing sessions as navigational and did not consider them in statistics dealing with reading volumes.

Fig. 3 shows the cumulative distribution of sessions per *CORE_USER* during the logging period. The median value of 305 sessions corresponds roughly to a daily usage pattern. However, 3.1% of Bookmate subscribers logged in more than 1,000 times during the 10 months.

DIGITAL READINGS BEHAVIOURS

This section describes the results of our analysis based on the four selected reading characteristics.

Reading Volume and Preferences

The reading volume analysis focuses on the absolute number of books readers consume during a certain period and whether book consumption is dependent on characteristics like gender, genre or length.

Number of Books

Figure 4 shows how many new books are opened by the Bookmate *CORE_USERS* in a month. The median is at 5.5 books per month, which already appears high, but there is a considerable number of users, who open up 10 books or more monthly (ca. 20%). This number is much higher than country average: according to VCIOM (the Russian Public Opinion Research Center), the average number of books read in three months is 6.26 in big cities (the numbers are somewhat higher for women and the elderly[15]). This seems to correlate with our initially stated assumption that readers on a streaming content delivery model will try out more books (not necessarily finish reading them) than an average reader would.

Genre and Gender Preferences

The reading preference analysis focuses on which books are accessed most often by the readers and whether these preferences differ by gender, length or genre of the book.

[15] http://wciom.ru/index.php?id=236\&uid=114843 (in Russian)

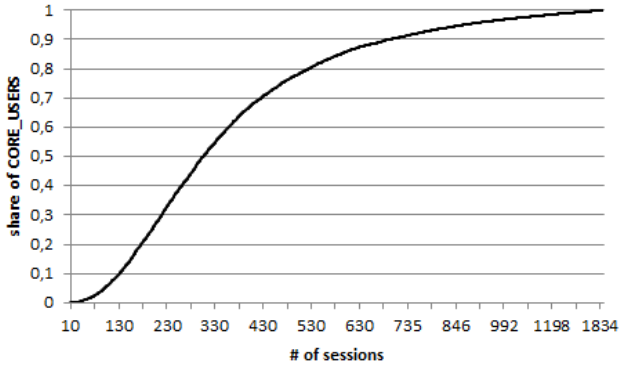

Figure 3. Cumulative distribution of reading sessions by users.

Figure 5 shows the popularity of genres based on the number of characters/books read by male/female users (note that the analysis is restricted by books with genre tags and users with specified gender).

The juxtaposition of books and characters read in each genre shows the advantage of using `low-level' log data: we can quantify not only genre preferences based on titles, but also engagement with specific genres. The figure reflects also indirectly average book length in different genres and completion rates (see also section Reading Style).

As we can see, fiction genres prevail – the total amount of fiction reading in characters is 4-5 times higher than non-fiction reading.

As could be expected, some genres are more popular among male readers (*Science Fiction & Fantasy, Business, Politics & Society, Technology & Science, History*), whereas others are preferred by female readers (*Love & Romance, Psychology, Art & Culture, Sports & Health*). Female users read notably more children books. There is no significant gender difference in *Modern Fiction, Crime & Thrillers, Mystery, Classics, Personal Growth*, and *Religion*.

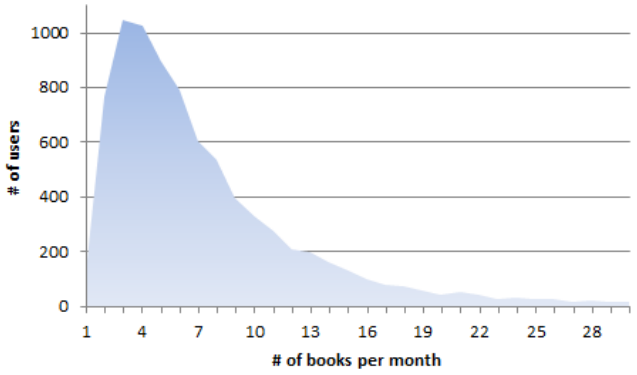

Figure 4. Distribution of users by the average number of books they open monthly. The figure represents the *normalized* number of unique books read by a user within the period. For example, if a reader was reading a single book during 10 months, it will result in 0.1 books per month.

Some books demonstrate distinct gender preferences. For example, while *Sad Cypress* by Agatha Christie and *Au Bonheur des Dames* by Emile Zola are almost exclusively read by women, there is 31% and 46% male readership of *Bridget Jones's Diary* by Helen Fielding and *Fifty Shades of Grey* by E. L. James, respectively.

An average male reader reads slightly more than a female; the difference is not statistically significant, however. There is also no difference in reading session length or access frequency for female and male readers. This contradicts the results of many studies stating that women read generally more[16]. We hypothesize that male Bookmate users are more engaged with the mobile application (i.e. read less complementary printed books) than female users.

4.2 Reading Schedule

The reading schedule analysis focuses on the times Bookmate subscribers prefer to read, whether readers have regular reading times and whether the schedule changes based on gender or genre.

Weekly & Daily Reading Patterns

Reading logs allow us to uncover reading schedule at different scales: hours, days, and weeks. Fig. 6 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

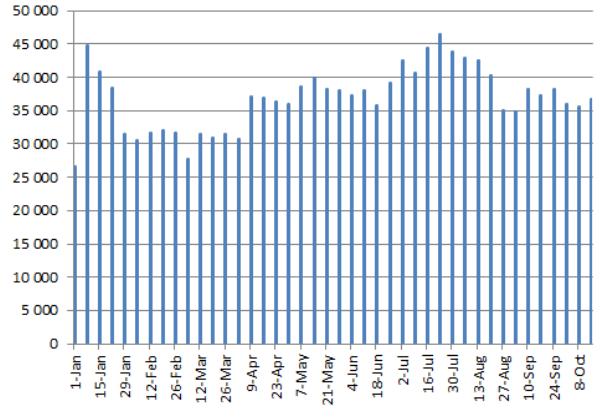

Figure 6. Average user weekly reading volume.

holidays (first decade of May). This indicates a 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 speculate that some Bookmate users prefer other leisure activities than reading on Saturdays[17].

There is no difference in weekly reading patterns in female and male users. There is also no marked preference for either fiction or non-fiction genres during weekdays or weekends.

Fig. 7 shows the relative distribution of reading activities for fiction and non-fiction genres throughout a day. It also 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 morning (10am–12pm), while fiction wins back afternoon and in the late night. During other times of the day fiction and non-fiction behavior is identical. There are only very subtle differences in the female/male reading schedules throughout the day (see Figure 8).

Reading Sessions

An average reading session is about 30 minutes long. Removing outliers (i.e. the sessions under 5 minutes and above 8 hours), the average reading session length becomes 41 minutes. Figures 9 and 10 show the distribution of gaps between reading sessions in hourly and daily ranges.

[16] See, for instance, the above cited VCIOM survey or a comprehensive report on reading in the USA (Bradshaw and Nichols, 2004).

[17] The survey of American reading habits (Bradshaw and Nichols, 2004) showed that frequent readers are more socially active than non-readers, visiting museums and attending shows, concerts, and sports events more often.

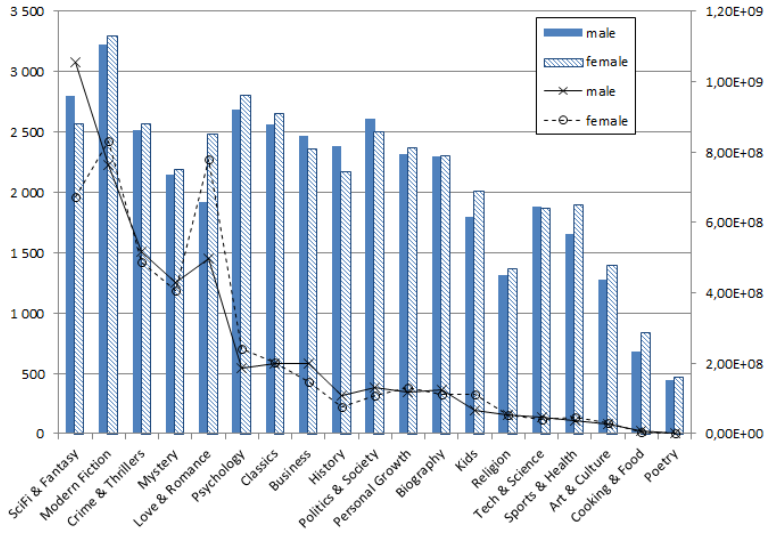

Figure 5. Genre popularity in female and male readers: number of books (bars) and characters read (lines); genres are ranked by the total characters read.

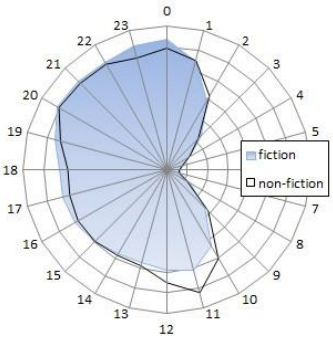

Figure 7. Relative volumes read by hours of the day for fiction vs. non-fiction.

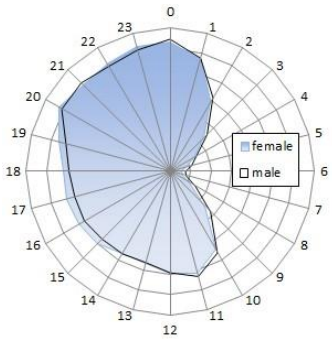

Figure 8. Amount of characters read by female and male users hours of the day.

Shorter gaps show the very frequent readers, who seem to access the Bookmate app whenever they have time or the inclination. Then, we can notice some increase in the range of 7–11 hours, which corresponds to two reading sittings a day (see also Figure 8). The next spike is around 24 hours (daily readings), followed by lower spikes corresponding to the whole number of days (see Figure 10).

Reading Speed

We view this log-based analysis of the reading speed as both a large-scale reading proficiency test (when focusing on the reading speed of users) and a readability study (when looking at how fast the books are read).

Figure 11 shows reading speed for Bookmate users in words per minute. The distribution is bell-shaped with mean around 150 words/min, which is the upper bound of the recommended reading speed for elementary school graduates.

Although we cannot distinguish between actual reading and navigational browsing in this data (which might skew the distribution), we believe the distribution reflects a realistic state for reading speeds.

The analysis of reading speeds per gender shows that women and men read at equal pace. Fig. 12 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 for adults appear in the ‘difficult’ sub spectrum. This suggests that these books are either read by the kids themselves or by parents aloud.

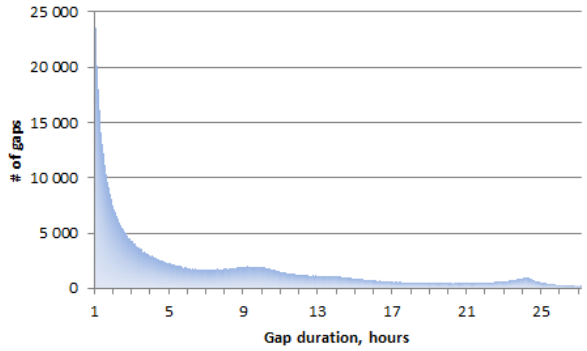

Figure 9. Gaps between reading sessions in hours.

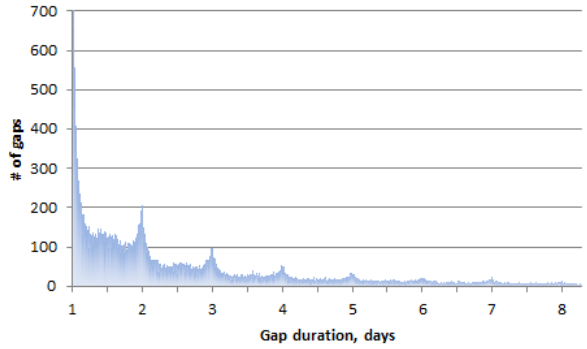

Figure 10. Gaps between reading sessions in days.

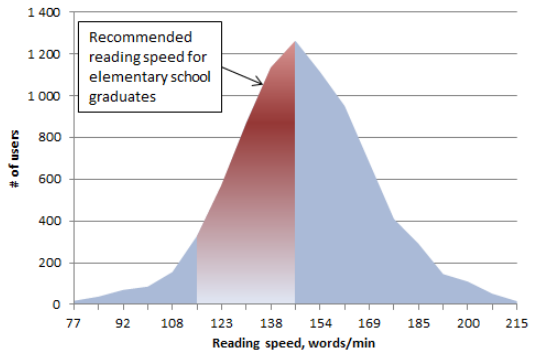

Figure 11. Users by their average reading speed (words per minute).

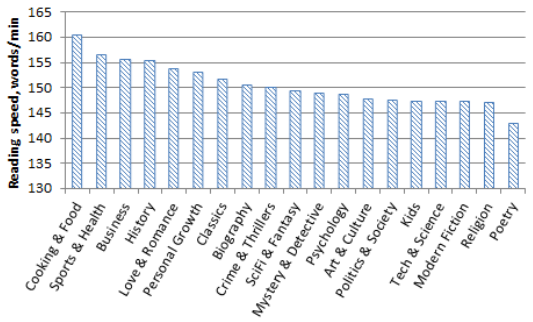

Figure 12. Reading speed averaged over genres.

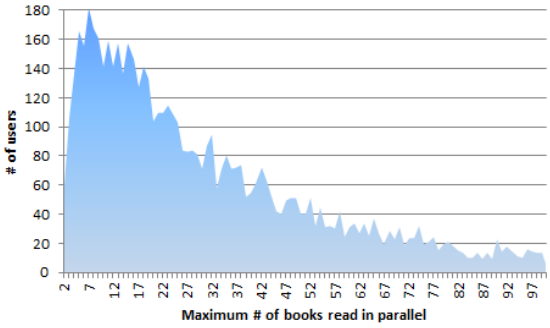

Figure 13. Users, who have `interleaving' book readings, i.e. read books interchangeably.

Reading Style

The reading style analysis focuses on the navigation patterns of readers within the books and in-between books. It looks at how and whether readers jump between different books (parallel and interleaving reading), which books are abandoned and when and whether this is dependent on gender, book length or genre.

Parallel and Interleaving Reading

As mentioned before, our hypothesis is that the easy access to a large collection tempts readers to read several books in parallel. We formalize parallel reading as follows: For each user we have a time-ordered sequence of reading sessions corresponding to different books:

$$b_{1,1}, b_{5,2}, b_{1,3}, \ldots, b_{i,k}, \ldots,$$

where $b_{i,k}$ is the ith book in the kth reading session. For every unique book that appears more than once in the sequence, we find the maximum span and calculate the number of different books within the span. Then, we find the maximum over all books read by the user.

We could observe parallel reading patterns in 6,579 (78.9%) of the *CORE_USERS*. Figure 13 shows the distribution of the maximum number of books read in parallel for these users. If we consider users who do not read in parallel (and disregard those 44 user outliers with more than 300 parallel books), the averaged maximum of books read in parallel is 32 and the median 17 books.

An analysis by gender shows that parallel reading behavior does not differ between men and women.

Book Completion and Abandonment

As we pointed out above, the low-level analysis of reading of a large user population is the most exciting opportunity that reading application logs provide. In particular, it is interesting to know what books are read to the end and what books are abandoned earlier. Again, we speculate that access to a large collection of ebooks promotes 'try-and-drop' reading patterns, especially in the streaming subscription model although direct comparisons with printed books are hardly feasible.

We consider a book abandoned if a user does not return to it within one month after last opening it (thus, a final reading

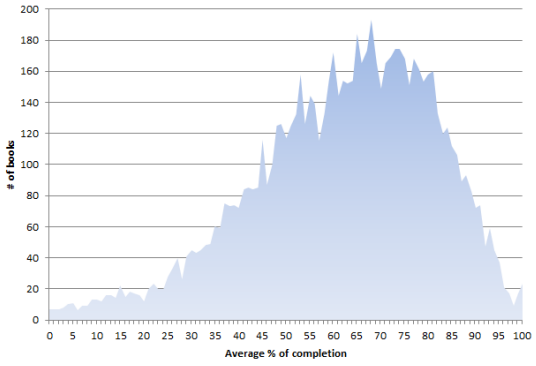

Figure 14. Numbers of books read until the specified percent of their length.

must take place not later than 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 characters 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 a completion rate above 90%. This score is similar to what the Kobo service reported for non-fiction books across different countries, but much lower for fiction books[18]. Since our number averages over all books, which contain more fiction books than non- fiction books, we could either assume that the Russian Bookmate readers lose interest faster than their e-reader counterparts in other countries based on nationality or – an explanation we find more likely – based on different subscription models. Kobo sells individual ebooks and does not offer the streaming content subscription model that Bookmate does.

Another figure on book completion can be found in the Goodreads survey: 38.1% of users reported that they always finish the book when started[19]. Bookmate users show a much lower persistence: only 111 (1.3%) of them read at least 90% of their books until 90% of their length and beyond.

In addition, we calculated the average point of abandonment for a subset of popular books (8,274 books with at least 10 readers), distributed as shown in the Fig. 14. When readers abandon books, they only abandon them after a thorough reading. 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, average 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. However, genre might play a role as was also confirmed by the Kobo survey mentioned previously.

[18] http://nyti.ms/1QFpcWz
[19] http://bit.ly/1RAM32Y

CONCLUSION AND FUTURE WORK

In this paper, we described results from an analysis of the Bookmate application log that corresponds to 10 months and 3 million reading sessions for about 8,000 of their regular users. We postulated that four variables would impact reading behavior: the reader, formal features of the book, the book content and the reading environment. Due to data constraints, we were able to study aspects from three of these four variables. We analyzed differences in reading behavior based on gender, book length and genre of the book. The main results of the presented analysis are following:

- Bookmate users' reading corresponds to a leisure-type activity: reading activity increases during evening and night and during holidays and vacation periods. A moderate decrease on Saturdays could be associated with alternative leisure activities practiced on weekends.
- Non-fiction is read more than fiction in the morning hours, while fiction dominates the night hours.
- Differences in male and female active digital readers can mainly be detected in genre preferences. Other studied aspects of reading behaviors are practically equivalent.
- Readers turn to their ebooks at least once a day on average.
- Ebook reading speeds correspond seemingly to those of printed books. There is no discernible difference between the speed needed for fiction and non-fiction books, only poetry is read much lower than average.
- The flat-rate subscription model seems to promote a ``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.
- Readers still show remarkable patience with a book: half of the books is only abandoned after two-thirds of the book's pages have already been read.

While previous ebook surveys and studies reported differences between printed and electronic resources as well as male and female reading behavior, our study cannot prove those significant variations.

It can be concluded that the reading logs are a valuable source of information about reading preferences, patterns, and behaviors, especially on the sub-book level.

Future work should especially analyze reading behavior with respect to the reading environment, i.e. the reading devices or reading location, as different behaviors can be expected. Due to data sparsity, this was not possible here. 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.

ACKNOWLEDGMENTS

We would like to 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

A. Adler, A. Gujar, B. L. Harrison, K. O'Hara, and A. Sellen. (1998). A diary study of work-related reading: Design implications for digital reading devices. In CHI '98, pages 241–248.

R. S. Akbar, H. A. Taqi, A. A. Dashti, and T. M. Sadeq. (2015). Does e-reading enhance reading fluency? English Language Teaching, 8(5):95–207.

T. Bradshaw and B. Nichols. (2004). Reading at risk: A survey of literary reading in America. Research division report# 46. Technical report, National Endowment for the Arts.

P. Braslavski, V. Petras, V. Likhosherstov, and M. Gäde. (2016) Ten months of digital reading: An exploratory log study. In TPDL'2016.

G. Buchanan, D. McKay, and J. Levitt (2015). Where my books go: Choice and place in digital reading. In JCDL '15, pages 17–26.

U. Eco and J.-C. Carrière. (2009) N'espérez pas vous débarrasser des livres. Grasset LGF, Paris.

A. Hinze, D. McKay, N. Vanderschantz, C. Timpany, and S. J. Cunningham. (2012). Book selection behavior in the physical library: implications for ebook collections. In JCDL'12, pages 305–314.

Y.-M. Huang, T.-H. Liang, C.-H. Chiu, et al. (2013). Gender differences in the reading of e-books: Investigating childrens attitudes, reading behaviors and outcomes. Educational Technology & Society, 16(4):97–110.

A. Hupfeld, A. Sellen, K. O'Hara, and T. Rodden. (2013). Leisure-based reading and the place of e-books in everyday life. In INTERACT'2013, pages 1–18.

J.-P. Hwang, Yueh-Min, et al. (2014). Investigating e-book reading patterns: A human factors perspective. In ICALT'2014, pages 104–108.

J. Y. Kim, H. Feild, and M. Cartright. (2012) Understanding book search behavior on the web. In CIKM'2012, pages 744–753.

H. Littlewood, A. Hinze, N. Vanderschantz, C. Timpany, and S. J. Cunningham. (2014). A log analysis study of 10 years of ebook consumption in academic library collections. In ICADL'2014, pages 171–181.

Z. Liu. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. J. Doc., 61(6):700–712.

Z. Liu and X. Huang. (2008) Gender differences in the online reading environment. J. Doc., 64(4):616–626.

C. Malama, M. Landoni, and R. Wilson. (2004). Fiction electronic books: A usability study. In ECDL'2004, pages 69–79.

C.C. Marshall. (2009). Reading and writing the electronic book. Synthesis lectures on information concepts, retrieval, and services, 1(1):1–185.

D. McKay. (2011). A jump to the left (and then a step to the right): Reading practices within academic ebooks. In OzCHI '11, pages 202–210.

D. McKay and G. Buchanan. (2016). You can check it out but it will never leave: Characterising ebook borrowing patterns. In CHIIR'2016, pages 203–212.

P. K. Murphy, J. F. Long, T. A. Holleran, and E. Esterly. (2003). Persuasion online or on paper: a new take on an old issue. Learning and Instruction, 13(5):511–532.

D. J. Oborne and D. Holton. (1988). Reading from screen versus paper: there is no difference. Int.J.Man-Mach.Stud., 28(1):1–9.

I. Rowlands, D. Nicholas, H. R. Jamali, and P. Huntington. (2007). What do faculty and students really think about e-books? Aslib Proceedings, 59(6):489–511.

A. Simpson, M. Walsh, and J. Rowsell. (2013). The digital reading path: researching modes and multidirectionality with ipads. Literacy, 47(3):123–130.

J. Staiger. (2012). How e-books are used: A literature review of the e-book studies conducted from 2006 to 2011. Reference and User Services Quarterly, 51(4):355–365.

J. C. Tucker. (2012). Ebook collection analysis: subject and publisher trends. Collection Building, 31(2):40–47.

T. M. Wagner, A. Benlian, T. Hess, et al. (2012). The role of product involvement in digital and physical reading-a comparative study of customer reviews of ebooks vs. printed books. In ECIS'2012.

C.Willis and M. Efron. (2013) Finding information in books: Characteristics of full-text searches in a collection of 10 million books. 50(1):1–10.

D. Zambarbieri and E. Carniglia. (2012). Eye movement analysis of reading from computer displays, ereaders and printed books. Ophthalmic and Physiological Optics, 32(5):390–396.

T. Zhang and X. Niu. (2015). Final report for 2015 er&l + ebsco library fellowship research project. In Libraries Reports, Paper 4.